PB82-190554

Determination of an Empirically Derived IP/TSP Relationship

PEDCo Environmental, Inc.
Durham, NC

Prepared for

Environmental Monitoring Systems Lab.
Research Triangle Park, NC

Apr 82

U.S. Department of Commerce
National Technical Information Service
NTIS.

PB82-190554

EPA-600/4-82-034
April 1982

DETERMINATION OF AN EMPIRICALLY DERIVED IP/TSP RELATIONSHIP

by

A. Carl Nelson, Jr.
Luke Wijnberg
PEDCo Environmental, Inc.
Durham, North Carolina 27701

Contract No. 68-02-3173
Task No. 27

Project Officers

Jack Suggs
EMSL, MD-56
Alice Gagnon
IERL, MD-63

ENVIRONMENTAL MONITORING SYSTEMS LABORATORY
U.S. ENVIRONMENTAL PROTECTION AGENCY
RESEARCH TRIANGLE PARK, NORTH CAROLINA 27711

August 1981

REPRODUCED BY
NATIONAL TECHNICAL
INFORMATION SERVICE
U.S. DEPARTMENT OF COMMERCE
SPRINGFIELD, VA 22161

TECHNICAL REPORT DATA
(Please read Instructions on the reverse before completing)

1. REPORT NO.	2.	3. RECIPIENT'S ACCESSION NO.
EPA-600/4-82-034	ORD Report	PB82 190554

4. TITLE AND SUBTITLE	5. REPORT DATE
Determination of an Empirically Derived IP/TSP Relationship	April 1982
	6. PERFORMING ORGANIZATION CODE
7. AUTHOR(S) A. Carl Nelson, Jr. Luke Wijnberg	8. PERFORMING ORGANIZATION REPORT NO.
9. PERFORMING ORGANIZATION NAME AND ADDRESS PEDCo Environmental, Inc. Durham, N.C . 27701	10. PROGRAM ELEMENT NO.
	11. CONTRACT/GRANT NO.
12. SPONSORING AGENCY NAME AND ADDRESS U. S. Environmental Protection Agency Enviornmental Monitoring Systems Laboratory Research Triangle Park, N.C. 27711	13. TYPE OF REPORT AND PERIOD COVERED Final - Aug. 1980-Feb. 81
	14. SPONSORING AGENCY CODE

15. SUPPLEMENTARY NOTES
Technical Assistance Document

16. ABSTRACT The primary objective of this study was to provide researchers with statistical methodology for comparing data on inhalable particulate (IP) and on the IP/TSP ratios from various sites, predicting IP concentration as a function of total suspended particulate (TSP) concentration, and detecting potential outliers. Eleven sites were selected for study based on the completeness of both IP and TSP data. The frequency distributions of IP and TSP data indicate that the lognormal distribution provides a better approximation than the normal distribution. No attempt is made to derive numerical results for all the available IP/TSP data because the 11 sites are not a random selection from a population of sites. The data from these 11 sites were compared graphically by means of box plots showing the dispersion of the date, the median value, and specified percentiles. One of the major results was the study of relationships between the measurements of IP and TSP. Two forms of this relationship were compared, a simple linear one and a first order approximation to a model relating the logarithms of IP and TSP. The linear relationship was considered to be satisfactory for most applications; however, the second model form indicates that the simple ratio prediction (i.e., using a constant ratio of IP/TSP over all values of TSP) may not be appropriate for all TSP. Simple statistical test procedures are given for comparing the average ratios at two or more sites (or for two or more years at a single test site), detecting outlying observations, and for selecting IP monitoring sites based on available TSP data and a proposed standard for IP.

17. KEY WORDS AND DOCUMENT ANALYSIS

a. DESCRIPTORS	b. IDENTIFIERS/OPEN ENDED TERMS	c. COSATI Field/Group
Inhalable particles IP/TSP Relationship (Model) Lognormal distribution Outlier tests	Inhalable articles Prediction relationships Transformed variables	

18. DISTRIBUTION STATEMENT	19. SECURITY CLASS *(This Report)*	21. NO. OF PAGES
Release to public	Unclassified	69
	20. SECURITY CLASS *(This page)*	22. PRICE
	Unclassified	

EPA Form 2220-1 (9-73)

DISCLAIMER

This report has been reviewed by the Environmental Monitoring Systems Laboratory, RTP, U.S. Environmental Protection Agency, and approved for publication. Approval does not signify that the contents necessarily reflect the views and policies of the U.S. Environmental Protection Agency, nor does mention of trade names or commercial products constitute endorsement or recommendation for use.

CONTENTS

	Page
Figures	v
Tables	vi
Acknowledgment	vii
Summary	viii
1. Introduction	1
2. Data Base for Analysis	3
Basic data sets used in analysis	3
Data set used in outlier analysis	6
3. Distribution of the Data	11
Tests for normality	11
Tests for lognormality	12
Tentative conclusions	12
Errors in both variables	14
4. Relationship of IP to TSP	15
5. Statistical Tests for Comparing Ratios and Regression Coefficients	28
Coventional methodology	28
Robust methodology	33
6. Determination of Outliers	34
Test using the IP/TSP ratio	34
Test using the bivariate normal (lognormal) assumption	37
Test using residuals (regression approach)	41
Conclusion	43
7. Procedure for Selection of IP Monitoring Sites Based on TSP Data	44
8. Conclusions	48
9. References	50

CONTENTS (continued)

		Page
Appendix A	Weighted regression analysis	52
Appendix B	Errors in both variables	53
Appendix C	Comparison of two measurement techniques	56
Appendix D	Model development	58

FIGURES

Number		Page
1	Sample Box Plot	5
2	Box Plots of the SSI/HVL Ratios	7
3	Box Plots of the TOT/HVL Ratios	8
4	Box Plots of the FNE/HVL Ratios	9
5	Box Plots of the CRS/HVL Ratios	10
6	Plot of TOT Residuals Versus HVL(TSP) for Model 1: $\hat{Y} = \hat{\beta}X$, weight = 1	18
7	Plot of TOT Residuals Versus HVL(TSP) for Model 1: $\hat{Y} = \hat{\beta}X$, weight = $1/X^2$	19
8	Plot of TOT Residuals Versus Log HVL(TSP) for Model 2: $\hat{Y} = \hat{\beta}_1 X + \hat{\beta}_2 X \log X$, weight = $1/X^2$	20
9	Probability Plot of the Residuals for Model 2: TOT Versus HVL for BIRM Data	21
10	Probability Plot of the Residuals for Model 2: FNE Versus HVL for BIRM Data	22
11	Probability Plot of the Residuals for Model 2: CRS Versus HVL for BIRM Data	23
12	SSI Versus HVL using Model 2 with PHILA2 Data	27
13	IP Versus TSP for Collocated SSI and Hi-vol Samplers	38
14	Log IP Versus Log TSP for Collocated SSI and HVL Samplers	40

TABLES

Number		Page
1	Number of Observations and High and Low Values of the Ratios	4
2	Results of Cramér-Von Mises Test (W^2) for Normality; Sample Size (n) and Significance Level	13
3	Results of Cramér-Von Mises Test (W^2) for Lognormality; Sample Size (n) and Significance Level	13
4	Results of Fitting Models 1 and 2 to 11 Basic Data Sets	16
5	Residual Mean Square (RESMSQ) and R^2 Values for Four Models	25
6	IP and TSP Data from PHILA 1 During 1979	30
7	IP and TSP Data from PHILA 2 During 1980	31
8	Inhalable Particulate (IP) and Total Suspended Particulate (TSP) Concentrations and Their Ratio	35

ACKNOWLEDGMENT

This report was prepared for the Environmental Monitoring Systems Laboratory of the U.S. Environmental Protection Agency. Mr. Jack Suggs was the Project Officer. PEDCo Environmental, Inc., appreciates the direction and report review provided. Mr. A. Carl Nelson, Jr. served as PEDCo's Project Manager. Dr. Louis Wijnberg performed a great deal of the data analysis and coauthored the report.

SUMMARY

The purpose of this report is to provide researchers with statistical methodology for comparing data on inhalable particulates (IP) from various sites, predicting IP concentration as a function of total suspended particulate (TSP) concentration, and detecting potential outliers. Data from 11 monitoring sites were used in the analysis because of the completeness of both IP and TSP measurements. Although the evidence is not conclusive, the frequency distributions of IP and TSP data from these 11 sites are better approximated by the lognormal distribution than by the normal distribution.

Two prediction relationships (models) are considered, linear models with both the original and the transformed variables (logarithms of the IP and TSP values). The log model led to the consideration of a two-term first-order approximation model in the original variables IP and TSP. This second model yielded a statistically significant improvement in the prediction relationship relative to the one-term linear model in about 25 percent of the cases. The one-term model, which predicts IP concentration as a constant fraction of the TSP concentration provides a good fit to the data and is therefore of considerable practical value.

If one compares the average IP/TSP ratio for different sites, the standard statistical t and F tests may be used, depending upon whether two or more sites are considered. If the number of measurements at each site is small, say less than 20, then the data should be transformed before performing the standard t and F tests. If IP and TSP are lognormally distributed then the ratio is also lognormally distributed. Applications of these tests are given in the report.

Various statistical test procedures can be used to determine outliers when both IP and TSP data are available. One procedure is to compare the largest (or smallest) IP/TSP ratio with the mean value and then divide by the standard deviation of the ratio. This test procedure may be performed with the logarithms of the variables or with the untransformed variables. Generally, however, the logarithm would be preferred. A second procedure is

to assume that IP and TSP data have a bivariate normal (or lognormal) distribution and to use a probability ellipse. A third procedure is to consider the residuals from a curve fitted by the method of least squares.

The report also presents a procedure for obtaining a confidence interval estimate of the mean IP/TSP ratio based on data from several sites, for estimating the relative precisions of two techniques for measuring IP concentration, and for selecting IP monitoring sites based on TSP data.

The 11 sites used in this study were selected because of data completeness and not as a representative sample of a specified population of sites. Hence, there is no attempt to provide numerical conclusions or parameter estimates concerning a population of measurements, except to illustrate a statistical methodology. The methodologies presented herein should have wide application to IP/TSP data.

SECTION 1

INTRODUCTION

The U.S. Environmental Protection Agency (EPA) is currently reviewing the ambient air quality standard for particulate matter. One revision being considered is the establishment of a standard for inhalable particulates (IP), which are defined as particles smaller than 15 μm aerodynamic diameter. There are numerous high-volume samplers for measuring total suspended particulate (TSP), but relatively few monitors (about 130 as of August 1981) for measuring IP. Thus, a review of the standard for TSP and hence for IP can be made more completely if the relationship between IP and TSP is better understood.

The primary purposes of this report are to investigate the distributions of IP, TSP, and the IP/TSP ratio, to study the relationship of IP to TSP, to compare the relationship for several sites, and to examine statistical tests for outliers.

Eleven sites were selected for investigation because of the availability of data on both IP and TSP. There was no attempt to use all available IP data because a separate study, for which there is an interim report,[1] is analyzing the entire IP data base and determining the means, standard deviations, and frequency distributions of the IP/TSP ratio for each site. The current investigation was limited to some of the statistical aspects of the data to avoid an overlap with the parallel study. The data base used in this investigation is described in Section 2.

Section 3 concerns the frequency distributions of the data, particularly, the application of normal and lognormal distributions. The effect of errors in both variables on least squares is also discussed in Section 3. Section 4 deals with the relationship between IP and TSP for single data sets, and Section 5 describes statistical tests for comparing these relationships for several data sets. Section 6 examines outlier tests. Section 7 discusses the problem of selecting IP monitoring sites based on TSP data and Section 8 lists the main conclusions.

This report contains four appendices. Appendix A, deals with weighted regression analyses; Appendix B, with errors in both variables; Appendix C, with the comparison of measurement techniques, and Appendix D, with the development of models for determining the IP/TSP ratio.

SECTION 2

DATA BASE FOR ANALYSIS

2.1 BASIC DATA SETS USED IN ANALYSIS

This report examines the relation between IP and TSP. The IP measurements are determined by two types of samplers: (1) a size-selective inlet (SSI) or modified hi-vol and (2) a dichotomous sampler.[2] The measurements by the dichotomous sampler are recorded as TOT (total), FNE (fine) and CRS (coarse), where the fine particles include those smaller than 2.5 μm aerodynamic diameter; the coarse particles, those smaller than 15 μm but greater than 2.5 μm; and thus the total includes both fine and coarse particles. The TSP measurements are determined by a hi-vol sampler (HVL).

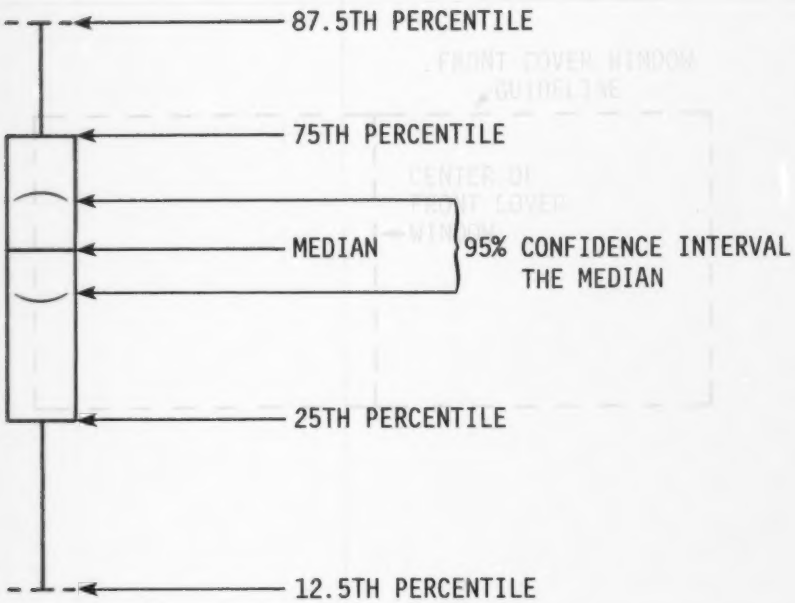
The data sets selected for analysis were those for which either or both of the ratios SSI/HVL and TOT/HVL could be computed, that is, on a given day HVL and SSI and/or TOT were measured. The number of observations in Table 1 are the total number of ratios available under the headings SSI/HVL, TOT/HVL, FNE/HVL, and CRS/HVL for a particular site. Storage and Retrieval of Aerometric Data (SAROAD) numbers are listed under each city. Multiple SAROAD numbers indicate that the data of more than one site are combined to make up the data set for that city. In the following cities two or more sites have been combined: Birmingham (BIRM), AL; Buffalo (BUFF), NY; Middletown (MIDD), OH; El Paso (ELPA), TX. The following cities are each represented by one site: Akron (AKRON), OH; Cincinnati (CINC), OH. The remaining cities, viz Los Angeles (LOSAN), CA and Philadelphia (PHILA), PA are each represented by multiple sites, treated as distinct data sets.

The highest and lowest ratios are also recorded on Table 1. Ratios in excess of unity are potential outliers. Very small ratios are also potential outliers as described in Section 6.

Ratio data are also presented in the form of box plots.[3] As shown in Figure 1, the top of the box (upper hinge) corresponds to the upper quartile; the bottom of the box corresponds to the lower quartile; the horizontal bar inside the box corresponds to the median; whiskers extend to the

TABLE 1. NUMBER OF OBSERVATIONS AND HIGH AND LOW VALUES OF RATIOS

	SSI/HVL (HI/LO)	TOT/HVL (HI/LO)	FNE/HVL (HI/LO)	CRS/HVL (HI/LO)
BIRM, AL 010380003 010380013 010380023	7 0.773 0.643	48 0.969 0.405	48 0.597 0.148	48 0.592 0.195
LOSAN 1, CA 054180103	13 0.838 0.370	35 0.997 0.444	35 0.655 0.102	35 0.485 0.187
LOSAN 2, CA 056535001	14 0.880 0.613	32 1.062 0.461	32 0.563 0.081	32 0.604 0.162
BUFF, NY 330660003 330660010	 no data	44 1.056 0.426	44 0.667 0.132	44 0.569 0.134
AKRON, OH 360060014	20 0.960 0.513	45 0.963 0.483	45 0.744 0.167	45 0.449 0.181
CINC, OH 361220020	10 0.824 0.524	36 1.032 0.450	36 0.720 0.208	36 0.497 0.123
MIDD, OH 364340005 364340006 364340007	49 0.952 0.508	29 0.759 0.470	29 0.463 0.107	29 0.528 0.140
PHILA 1, PA 397140003	108 0.993 0.337	26 1.269 0.527	26 0.663 0.227	26 0.820 0.168
PHILA 2, PA 397140019	80 1.054 0.408	17 1.152 0.144	17 0.610 0.068	17 0.599 0.077
PHILA 3, PA 397140038	89 1.015 0.468	16 0.966 0.620	16 0.657 0.365	16 0.469 0.222
ELPA, TX 451700002 451700004	13 0.981 0.453	34 1.367 0.282	34 0.551 0.025	34 0.816 0.053



THE WIDTH OF THE BOX IS PROPORTIONAL TO THE SQUARE ROOT OF THE SAMPLE SIZE.

Figure 1. Sample box plot.

upper and lower octiles, respectively. The parentheses define a 95% confidence interval around the median.[3,4] Figures 2 through 5 are box plots of the SSI/HVL, TOT/HVL, FNE/HVL, and CRS/HVL ratios, respectively. If the confidence intervals for two boxes do not overlap, we are highly confident that the corresponding population medians are different.[3,4] The widths of the boxes are proportional to the square root of the number of observations.

Throughout the investigation, the hi-vol (HVL) measurement has been assumed to be the carrier (independent or predictor variable) in a regression model, and thus, the denominator in ratio estimates. The other variables have been taken to be response variables and thus, numerators in ratio estimates. The four response variables are correlated, and their correlation has been properly attended to by using a multivariate linear model.

As shown by the box plots, the data sets can usually be subdivided into two or more groups, within which the medians are not significantly different. For example, box plots of the TOT/HVL ratio indicate that the data sets can be partitioned into two groups, one including (BIRM, LOSAN1, LOSAN2, BUFF, MIDD, PHILA2, ELPA) and the other including (AKRON, CINC, PHILA1, PHILA3). Within each group the medians are not significantly different; the medians of one group, however, tend to be significantly different from the medians of the second group.

2.2 DATA SET USED IN OUTLIER ANALYSIS

One data set used in subsequent analysis, particularly for outlier tests in Section 6, is taken from a previous PEDCo study.[5] The data are IP measurements made by an SSI sampler and TSP measurements by a collocated HVL sampler. Data on both IP and TSP are available for 37 days, with 1 day as a potential outlier as described in Section 6. PEDCo monitors were located at about 0.5 km south of a quarry and the average IP/TSP ratio was higher than that reported by several EPA monitoring sites.[5] The relatively high number of small particulates at a distance of 0.5 km from the quarry was a surprise finding for fugitive dust from quarries. Further discussion is in the PEDCo report. These data are used in this report for outlier analysis and not for comparison with other data, although Section 6 provides tests for making comparisons.

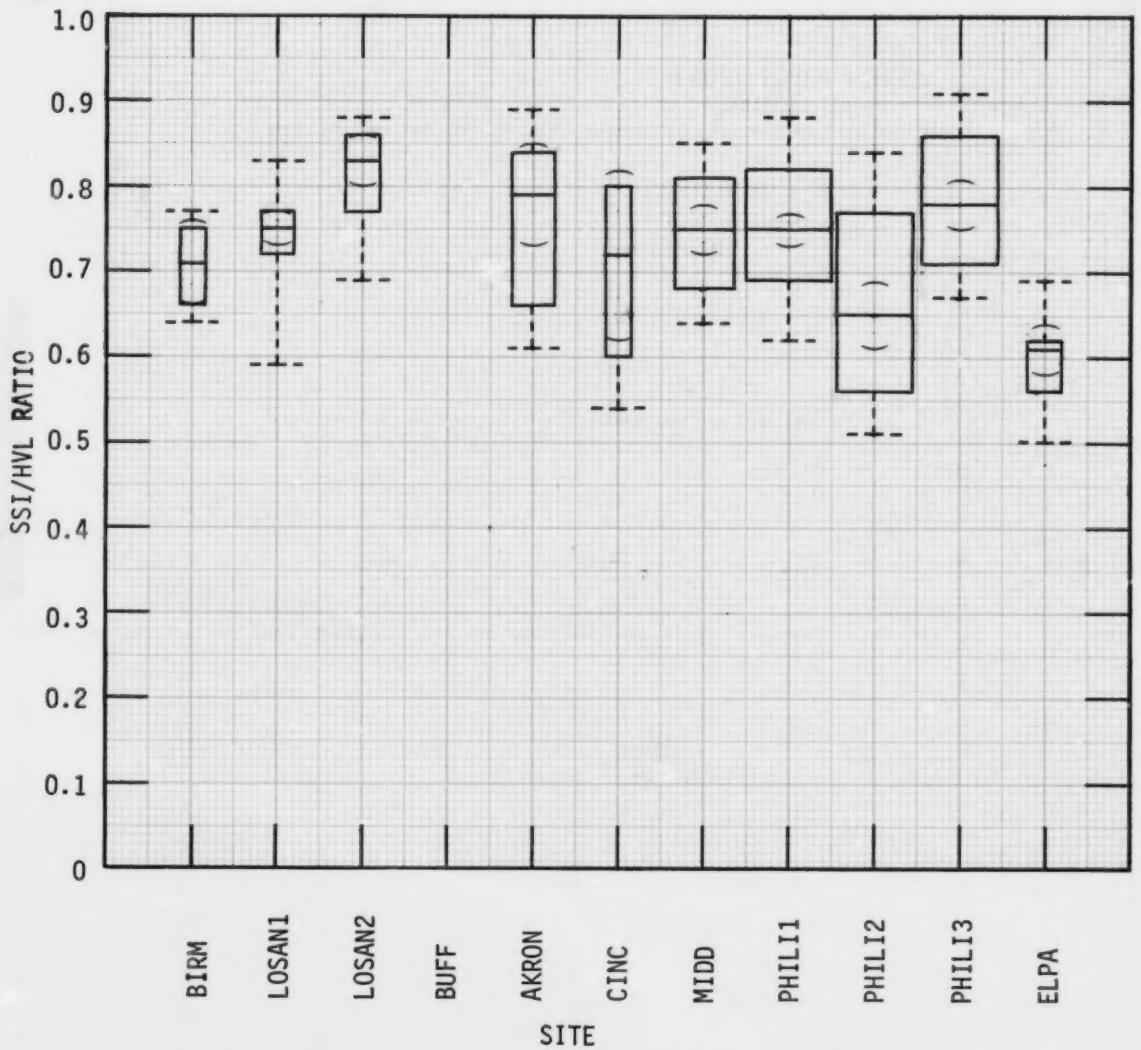


Figure 2. Box plots of the SSI/HVL ratios.

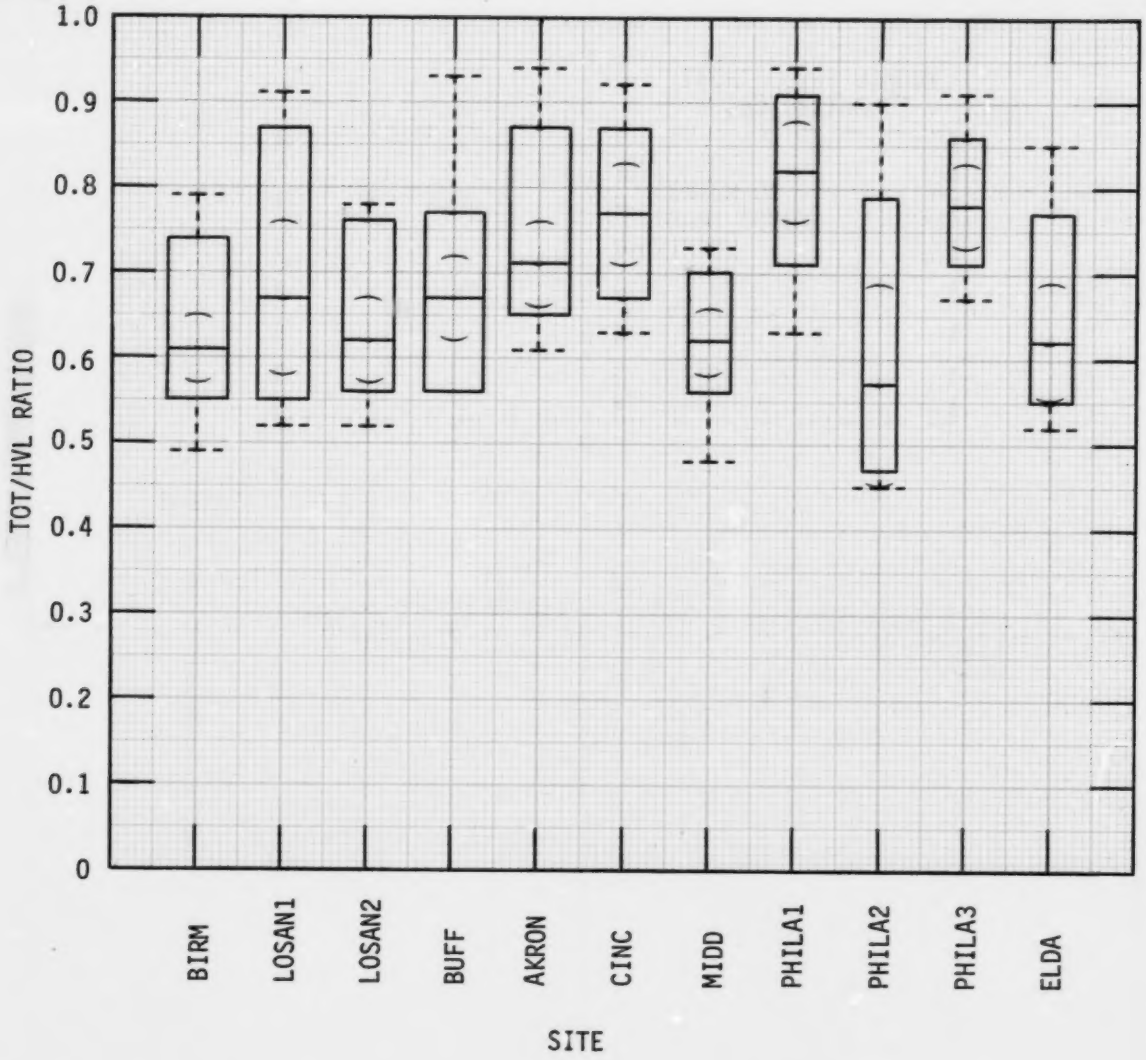


Figure 3. Box plots of the TOT/HVL ratios.

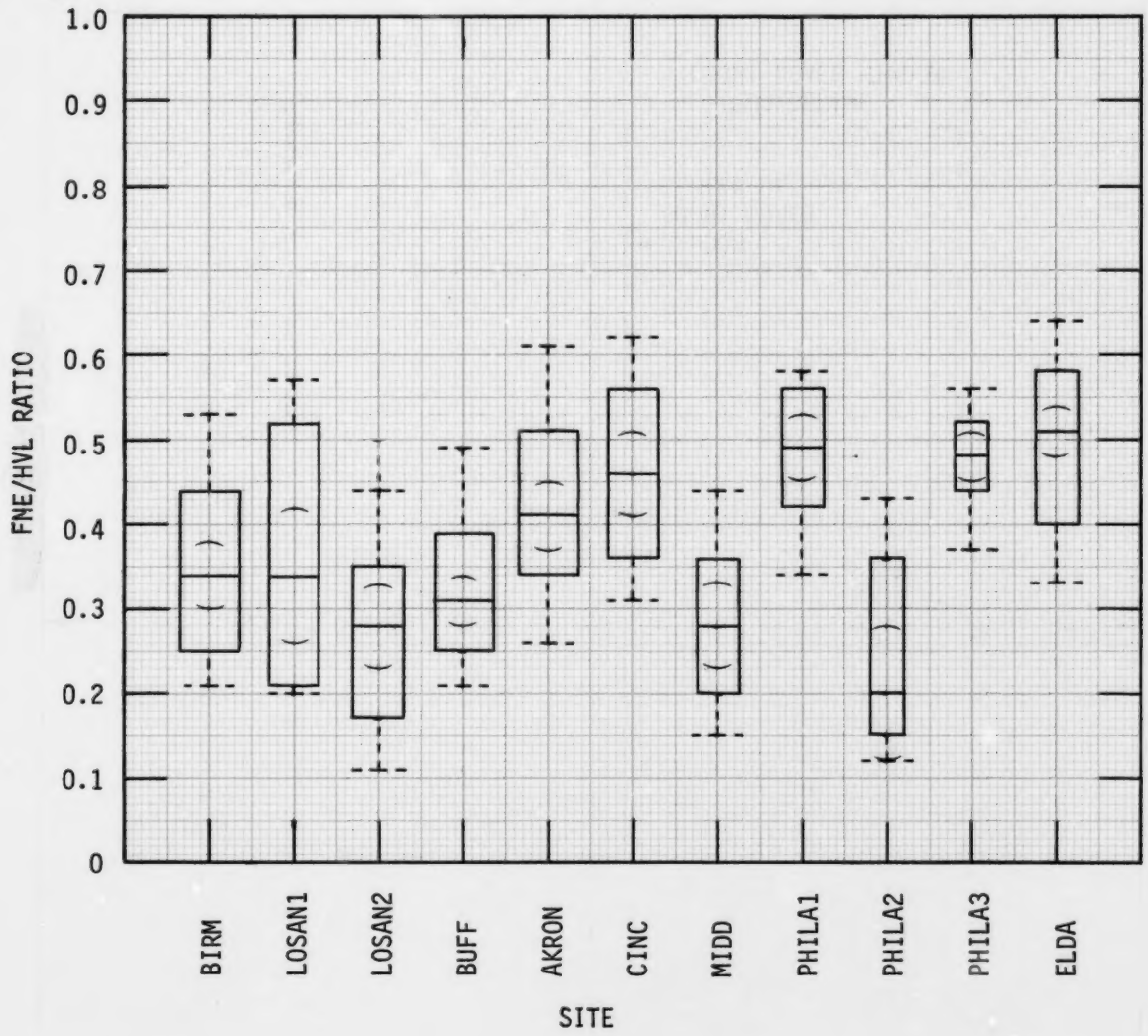


Figure 4. Box plots of the FNE/HVL ratios.

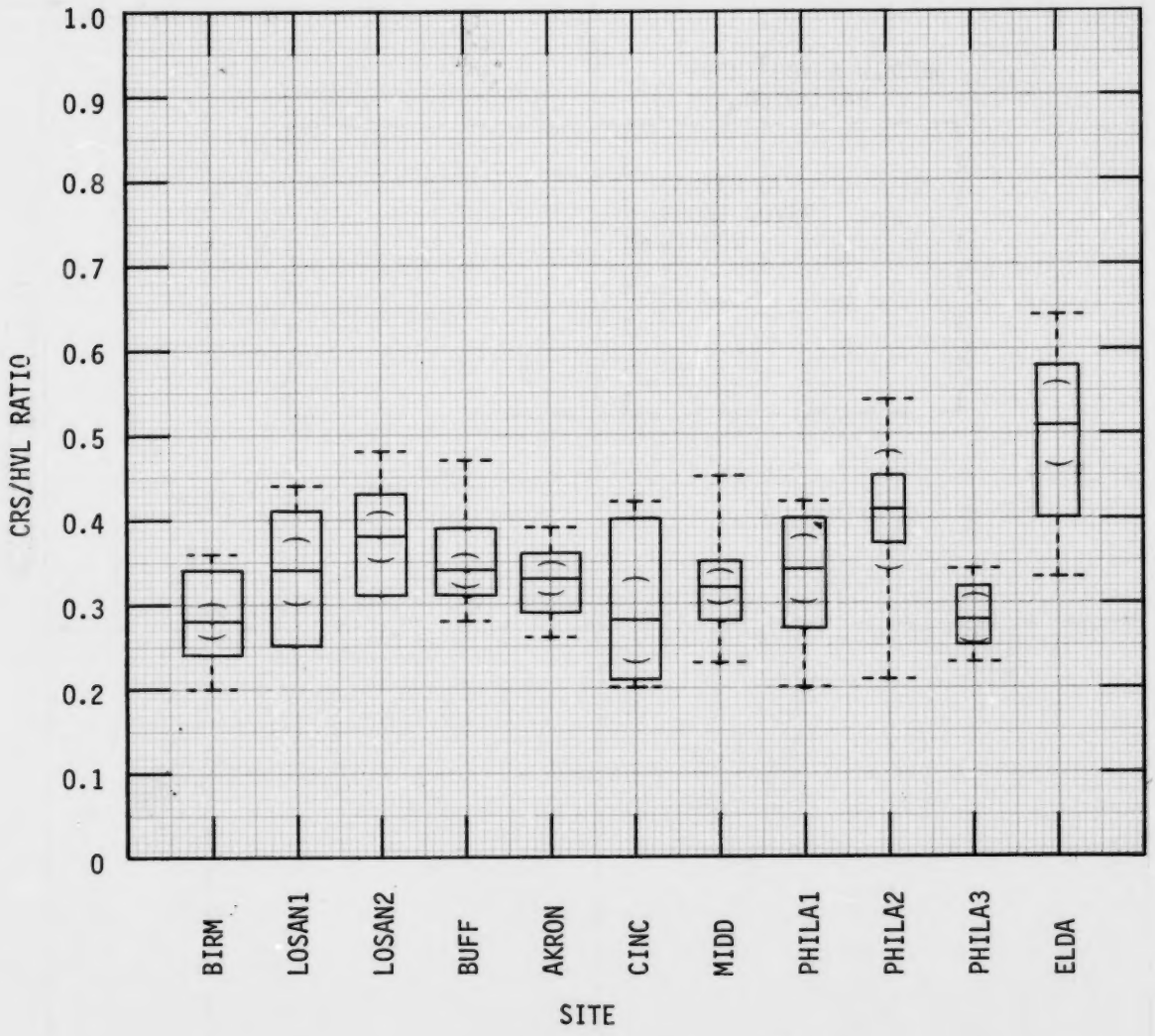


Figure 5. Box plots of the CRS/HVL ratios.

SECTION 3

DISTRIBUTION OF THE DATA

3.1 TESTS FOR NORMALITY

We have chosen three methods to test for normality. Two of these are recommended by Stephens:[6] the Cramér-von Mises statistic (W^2) and the Anderson-Darling statistic (A^2). They are given by

$$W^2 = n^{-1} \sum_{i=1}^{n} [F(X_i) - F_n(X_i)]^2$$

and

$$A^2 = -n^{-1} \sum_{i=1}^{n} (2i - 1)\{\log F(X_i) + \log [1 - F(X_{n + 1 - i})]\} - n,$$

where $F(X)$ is the (continuous) distribution function and $F_n(X)$ is the empirical distribution function. Percentage points for both statistics are available[6] and are valid when distributional parameters must be computed from the sample.

The third method is the probability plot as described in the Biomedical Computer Programs (BMDP)-79 P-Series Manual.[7] Observed data values are ordered from smallest to largest and then plotted along the horizontal axis. The vertical axis corresponds to the expected normal value based on the rank of the observation. Let $X_{(1)}$, $X_{(2)}$, ... $X_{(n)}$ represent the data values after they have been ordered from smallest to largest. The subscript (j) is the rank order of the observation. If n is the total number of observations, the vertical plotting position corresponds to the expected normal value for the relative rank (j out of n) of the observation. The expected normal value is estimated by

$$\Phi^{-1}[(3j - 1)/(3n + 1)],$$

which is the standard normal value corresponding to the probability $(3j - 1)/(3n + 1)$ and ϕ denotes the standard normal frequency function. If the data are from a normal distribution they tend to fall on a straight line.

Each measurement category (i.e., HVL, SSI, TOT, FNE and CRS) in the 11 basic data sets has been tested for normality by W^2 and A^2. The A^2 results are almost identical to the W^2 results, hence, only W^2 results are given in the Table 2. The number of measurements (n) and the level of significance (sign) are given in this table for each W^2 result.

3.2 TESTS FOR LOGNORMALITY

A subset of the data sets, (i.e., the AKRON, PHILA1, PHILA2, and PHILA3 sets) has been tested for lognormality by W^2 and A^2. That is, the observations were transformed by taking natural logarithms and then tested for normality. The lognormality results for W^2 are presented in Table 3.

3.3 TENTATIVE CONCLUSIONS

Tests for normality indicated a large proportion of significant results (i.e., 33 cases of the 54 cases were significant at the 0.10, 0.05, 0.025, or 0.01 level). On the other hand, tests for lognormality showed that only 9 of the 54 cases were significant, with 5 of these 9 at the 0.10 level. Hence, the lognormal distribution appears to provide a better approximation than the normal distribution; although, the evidence is not conclusive.

Other skewed distributions were not investigated because: (1) the lognormal distribution has been extensively used in fitting TSP data, and (2) the lognormal distribution is easy to work with mathematically. Furthermore, it is inferred that the lognormal is generally better than the normal distribution in fitting these data but not necessarily the best distribution among all possible distributions.

In the following sections the lognormal distribution is assumed to be more accurate for almost all applications; however, some results are discussed for both the normal and lognormal distributions. This assumption influences selection of the regressions models relating IP to TSP (or HVL), the tests for comparing average ratios for different sites, and the tests for outliers.

TABLE 2. RESULTS OF CRAMER-VON MISES TEST (W^2) FOR NORMALITY; SAMPLE SIZE (n) AND SIGNIFICANCE LEVEL*

	HVL			SSI			TOT			FNE			CRS		
Site(s)	W^2	n	Sign	W^2	n	Sign	W^2	n	Sign	W^2	n	Sign	W^2	n	Sign
BIRM	0.238	55	d	0.050	7	ns	0.091	48	ns	0.125	48	a	0.140	48	b
LOSAN1	0.123	35	a	0.039	13	ns	0.065	35	ns	0.145	35	b	0.068	35	ns
LOSAN2	0.067	35	ns	0.108	14	a	0.141	32	b	0.277	32	d	0.103	32	ns
BUFF	0.208	44	d	No data			0.123	44	a	0.162	44	c	0.151	44	c
AKRON	0.139	48	b	0.065	20	ns	0.089	45	ns	0.088	45	ns	0.109	45	a
CINC	0.058	37	ns	0.051	10	ns	0.050	36	ns	0.100	36	ns	0.083	36	ns
MIDD	0.095	49	ns	0.074	49	ns	0.068	29	ns	0.142	29	b	0.084	29	ns
PHILA1	0.440	114	d	0.278	26	d	0.128	26	b	0.105	26	a	0.104	26	ns
PHILA2	0.644	114	d	0.649	80	d	0.116	17	a	0.324	17	d	0.125	17	a
PHILA3	0.366	111	d	0.384	89	d	0.264	16	d	0.368	16	d	0.123	16	a
ELPA	0.230	41	d	0.205	13	d	0.105	34	a	0.215	34	d	0.103	34	ns

TABLE 3. RESULTS OF CRAMER-VON MISES TEST (W^2) FOR LOGNORMALITY; SAMPLE SIZE (n) AND SIGNIFICANCE LEVEL*

	HVL			SSI			TOT			FNE			CRS		
Site(s)	W^2	n	Sign	W^2	n	Sign	W^2	n	Sign	W^2	n	Sign	W^2	n	Sign
BIRM	0.602	55	ns	0.461	7	ns	0.061	48	ns	0.051	48	ns	0.070	48	ns
LOSAN1	0.041	35	ns	0.082	13	ns	0.030	35	ns	0.031	35	ns	0.056	35	ns
LOSAN2	0.050	35	ns	0.054	14	ns	0.038	32	ns	0.063	32	ns	0.081	32	ns
BUFF	0.042	44	ns	No data			0.033	44	ns	0.034	44	ns	0.069	44	ns
AKRON	0.080	48	ns	0.026	20	ns	0.042	45	ns	0.039	45	ns	0.068	45	ns
CINC	0.105	37	a	0.103	10	ns	0.122	36	a	0.066	36	ns	0.128	36	b
MIDD	0.040	49	ns	0.046	49	ns	0.072	29	ns	0.033	29	ns	0.052	29	ns
PHILA1	0.114	114	a	0.065	108	ns	0.047	26	ns	0.036	26	ns	0.027	26	ns
PHILA2	0.063	103	a	0.115	80	a	0.031	17	ns	0.063	17	ns	0.030	17	ns
PHILA3	0.366	106	d	0.050	83	ns	0.083	16	ns	0.136	16	b	0.031	16	ns
ELPA	0.094	41	ns	0.074	13	ns	0.086	34	ns	0.077	34	ns	0.156	34	c

ns = not significant (probability is >0.10).
a = significance at the 0.10 level.
b = significance at the 0.05 level.
c = significance at the 0.025 level.
d = significance at the 0.01 level.

3.4 ERRORS IN BOTH VARIABLES

One of the fundamental assumptions in regression analyses is that the "independent" variables are measured without error. Thus, if there is a measurement error in the X variable (in this case the HVL measurement), then the estimate of ratios involving X and estimates of regression coefficients using the standard techniques may be biased. Two basic references discuss statistical procedures for use when both the X and the Y variables are subject to measurement error and when a linear relationship is used.[8,9] Kunitomo obtains approximate distributions of the maximum-likelihood estimator ($\hat{\beta}_M$) and the ordinary least-squares estimator ($\hat{\beta}_L$) subject to the condition of errors in both variables. "On the whole, the cumulative distribution function (cdf) of $\hat{\beta}_L$ is biased downward and is more concentrated around $[c/(1 + c)]\ \beta$ than the cdf of $\hat{\beta}_M$ for a relatively large sample size.[9] The value c is the ratio of the variance of the true values of the x's by the variance of the errors in measurements of the individual x's. (Note: In Appendix B the true values are denoted by X, the observed values as x, and the errors in observations by u. Hence, in this notation $c = \sigma^2(X)/\sigma^2(u)$.) Hence, if c were 9, then the bias in estimating the slope (β) by the ordinary least squares method would be 0.1 β. In practical terms, if the data are sufficiently dispersed relative to the error in measurement then the effect of the error on the least-squares estimator is insignificant.

Appendix B shows the effect of errors in X measurement on estimates of the true slope. In summary, the ratio estimate $(\frac{1}{n} \Sigma \frac{Y}{X})$ is biased less than 1 percent because of the assumed error of 5 percent (relative standard deviation of HVL data) in X measurement.

SECTION 4

RELATIONSHIP OF IP TO TSP

Scatterplots of the response variables Y (SSI, TOT, FNE and CRS) versus X (HVL) suggest a linear relation.[1] In addition, these plots suggest that the residual variance of the Y response variables will increase as X increases. Distributional tests on the variables indicate that a log transformation should be considered. This transformation also has the advantage of stabilizing the variance of Y for a given X, where the variance is proportional to X^2. For these reasons a preliminary study was made by fitting a simple regression of log Y on log X for six sites (the three Philadelphia sites, the Akron site, and two Middletown sites). The results suggested that the simple linear model,

$$\text{Model 1:} \quad Y = \beta X + U_1, \tag{1}$$

is usually adequate, however, the model

$$\text{Model 2:} \quad Y = \beta_1 X + \beta_2 X \log X + U_2 \tag{2}$$

sometimes yields a significant improvement in the prediction. The β, β_1, and β_2 are the regression coefficients and U_1 and U_2 are vectors of the errors (residuals) in Y for given X. The second model is derived in Appendix D and it includes the first model as a special case. The second model is also a first order approximation of the model, $\log Y = \alpha + \beta \log X$, the log model. The coefficients of Equations (1) and (2) are estimated by multiple linear regression with weights proportional to $1/X^2$. (It is not necessary that the weights sum to unity [i.e., $\Sigma\omega_i = 1$] because the computational scheme automatically takes care of this fact.) The difference between Models 1 and 2 is the curvilinear term X log X. In a portion of the cases tested, the addition of this term does not improve prediction and hence the simple linear model emerges naturally.

Models 1 and 2 were fitted to each of the 11 basic data sets and the results are in Table 4. As explained in Appendix A, the weighted regression estimate of β in Model 1 is the average of the ratios ($\bar{R}$), that is,

TABLE 4. RESULTS OF FITTING MODELS 1 AND 2 TO 11 BASIC DATA SETS

		$\hat{\beta}$	$\hat{\beta}^1$	Sign of $\hat{\beta}^1$	$\hat{\beta}^2$	Sign of $\hat{\beta}^2$	R^2	n
	SSI	0.711	0.530	c	0.038	ns	0.997	7
BIRM	TOT	0.641	1.203	d	0.135	d	0.974	48
AL	FNE	0.355	1.012	d	0.158	d	0.923	48
	CRS	0.286	0.195	a	0.022	ns	0.940	48
	SSI	0.718	COL		0.170	d	0.977	13
LOSAN1	TOT	0.708	0.708	d	COL		0.943	35
CA	FNE	0.363	0.363	d	COL		0.838	35
	CRS	0.344	0.344	d	COL		0.940	35
	SSI	0.811	0.810	d	COL		0.992	14
LOSAN2	TOT	0.674	0.506	c	0.223	ns	0.964	32
CA	FNE	0.278	0.035	ns	0.012	ns	0.822	32
	CRS	0.404	0.298	a	0.022	ns	0.937	32
	SSI		No data					0
BUFF	TOT	0.692	0.692	d	COL		0.952	44
NY	FNE	0.335	0.335	d	COL		0.880	44
	CRS	0.357	0.357	d	COL		0.946	44
	SSI	0.761	0.761	d	COL		0.978	20
AKRON	TOT	0.748	0.748	d	COL		0.967	45
OH	FNE	0.422	0.422	d	COL		0.904	45
	CRS	0.326	0.326	d	COL		0.968	45
	SSI	0.691	0.594	ns	0.024	ns	0.979	10
CINC	TOT	0.467	0.480	c	0.003	ns	0.972	36
OH	FNE	0.467	0.480	c	0.003	ns	0.926	36
	CRS	0.301	0.584	d	0.072	b	0.913	36
	SSI	0.748	1.121	d	0.089	d	0.987	49
MIDD	TOT	0.609	0.723	d	0.026	ns	0.978	29
OH	FNE	0.285	0.811	d	0.121	d	0.910	29
	CRS	0.324	0.089	ns	0.949	d	0.944	29
	SSI	0.749	0.749	d	COL		0.978	108
PHILA1	TOT	0.812	1.622	d	0.201	d	0.976	20
PA	FNE	0.474	0.889	d	0.103	c	0.965	20
	CRS	0.357	0.793	d	0.109	a	0.878	20
	SSI	0.660	1.213	d	0.121	d	0.967	114
PHILA2	TOT	0.625	1.603	c	0.215	ns	0.886	17
PA	FNE	0.273	1.040	c	0.168	a	0.800	17
	CRS	0.352	0.563	ns	0.046	ns	0.864	17
	SSI	0.782	0.639	d	0.035	ns	0.981	89
PHILA3	TOT	0.780	1.091	d	0.078	a	0.990	16
PA	FNE	0.485	0.733	d	0.063	a	0.982	16
	CRS	0.296	0.358	c	0.016	ns	0.957	16
	SSI	0.641	COL		0.139	d	0.982	13
ELPA	TOT	0.671	1.319	d	0.147	d	0.931	34
TX	FNE	0.178	0.623	d	0.101	d	0.794	34
	CRS	0.493	0.695	d	0.046	ns	0.924	34

ns = not significant (probability is >0.10).
a = significance at the 0.10 level.
b = significance at the 0.05 level.
c = significance at the 0.025 level.
d = significance at the 0.01 level.
COL = colinearity.

$$\bar{R} = \frac{1}{n} \sum_{i=1}^{n} (Y_i/X_i) \qquad (3)$$

Table 4 lists the estimated coefficients ($\hat{\beta}_1$ and $\hat{\beta}_2$), the significance of the individual coefficients in fitting Model 2, and the square of the multiple correlation coefficient (R^2) obtained by use of BMDP program P6R (a multivariate, weighted, multiple linear regression).[7]

In the case of colinearity (COL) the second model usually reduces to Model 1. In two cases of COL (LOSAN 1, Y = SSI; ELPA, Y = SSI), the algorithm led to the result

$$\hat{Y} = \hat{\beta}_2 X \log X.$$

One can detect an improvement in the symmetry of the residuals about the origin based on the use of weight proportional to $(HVL)^{-2}$ or $1/X^2$. This observation supports the choice of weights. Examples are given in Figures 6 through 8. The residuals in Figures 7 and 8 are more evenly distributed about zero than those in Figure 6; in particular, the residuals in Figure 8 (Model 2) indicate no trend with the independent variable (log HVL). Thus, in some cases a real improvement in the fit results from using Model 2 in preference to Model 1.

If the two terms X, and X log X are computationally colinear (i.e., if they are colinear within the tolerance of the computer algorithm), then several options are available:

1. Improve the tolerance, (e.g., by use of double precision, which is available in BMDP·P9R.
2. Use principal component regression.[10]
3. Use ridge regression.[10]
4. Discard one of the two terms. The BMDP·P6R discards one of the computationally colinear terms.

The measure of the adequacy of fit is R^2. An R^2 value close to unity may indicate a good fit, but not necessarily so.[10] Visual examination of the probability plots of the residuals in Figures 9, 10, and 11 suggests

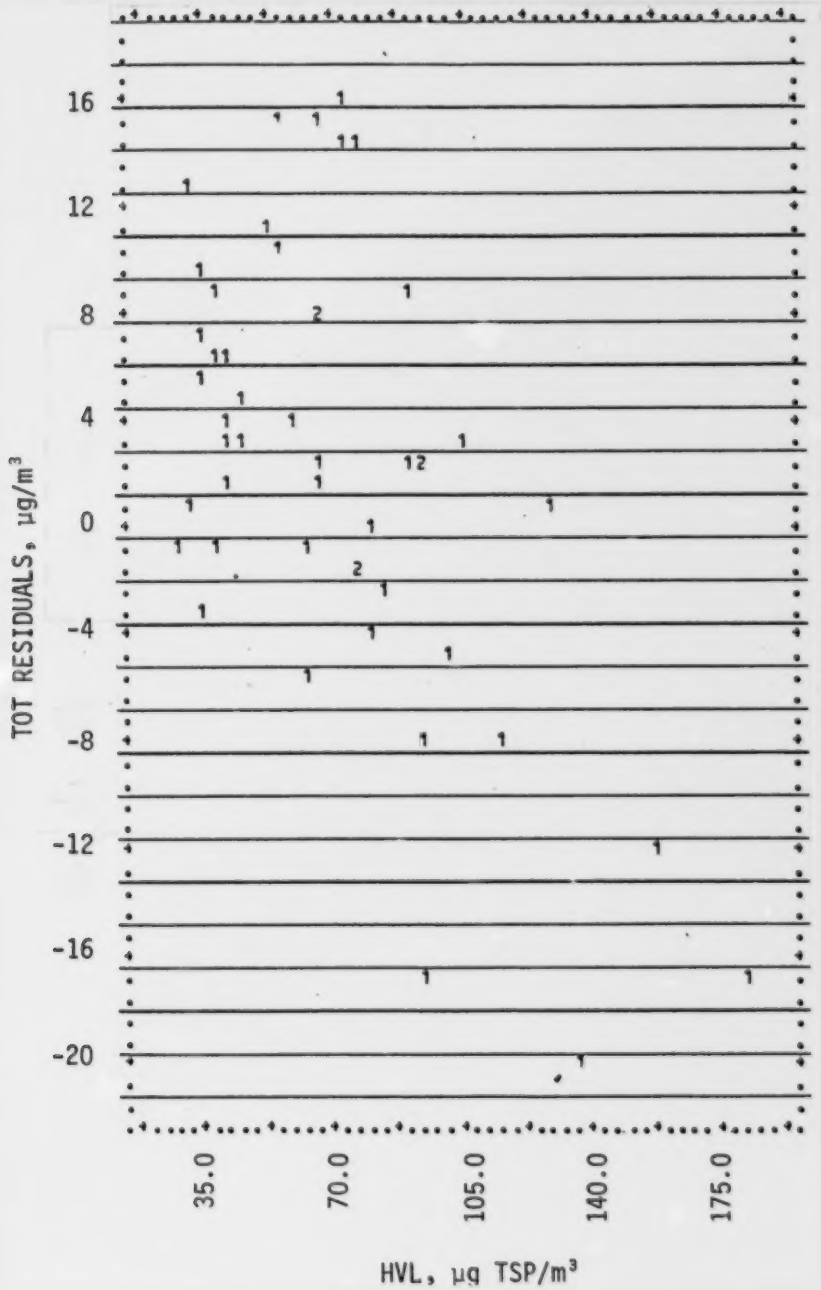


Figure 6. Plot of TOT residuals versus HVL(TSP) for Model 1: $\hat{Y} = \hat{\beta}X$, wt = 1.

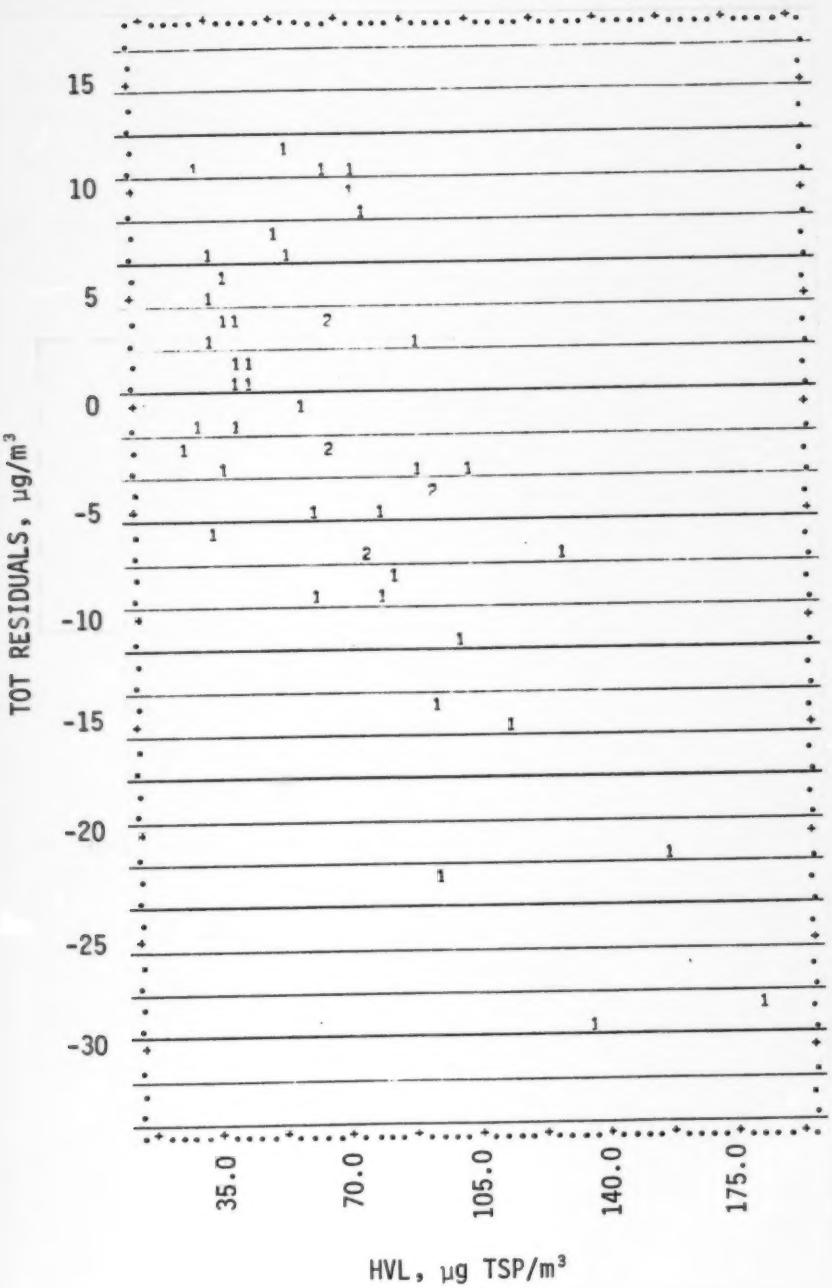


Figure 7. Plot of TOT residuals versus HVL(TSP) for Model 1: $\hat{Y} = \hat{\beta}X$, wt = $1/X^2$.

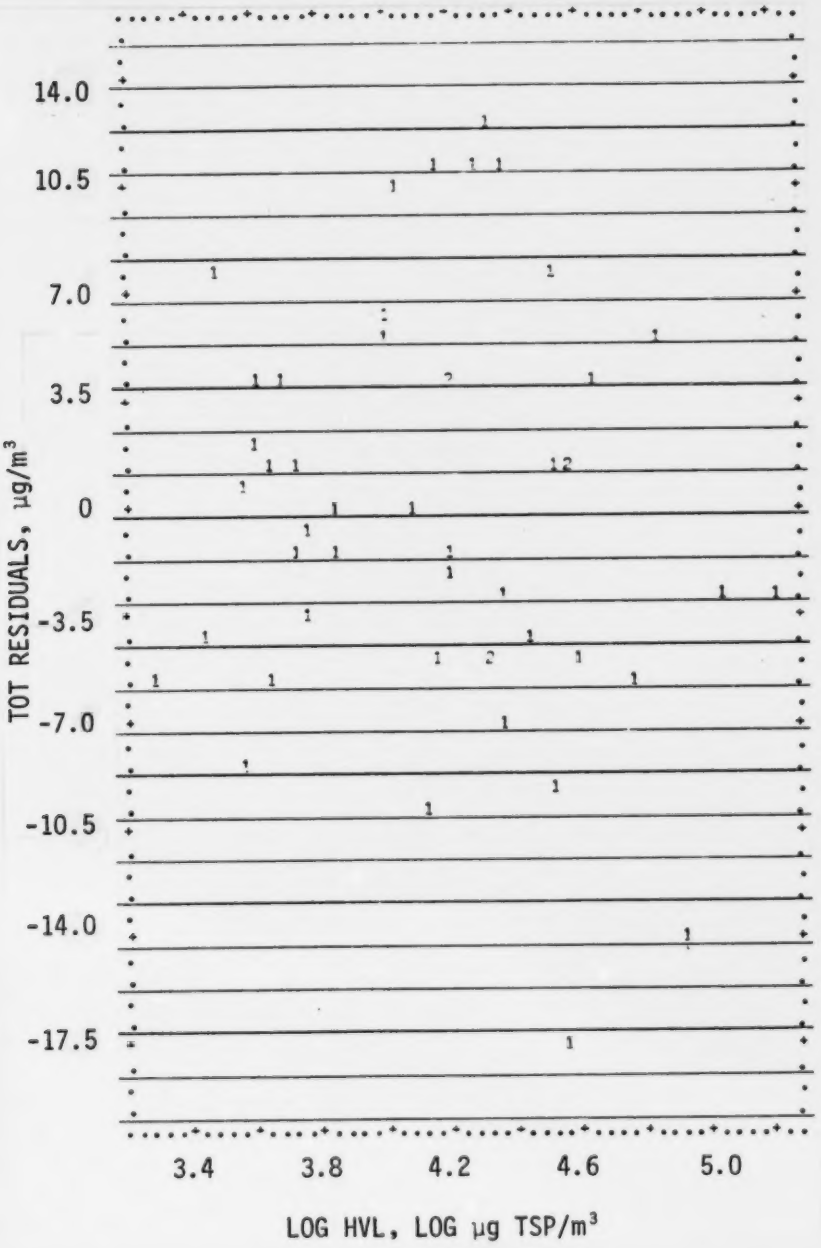


Figure 8. Plot of TOT residuals versus LOG HVL(TSP) for Model 2: $\hat{Y} = \hat{\beta}_1 X + \hat{\beta}_2 X \log X$, wt = $1/X^2$.

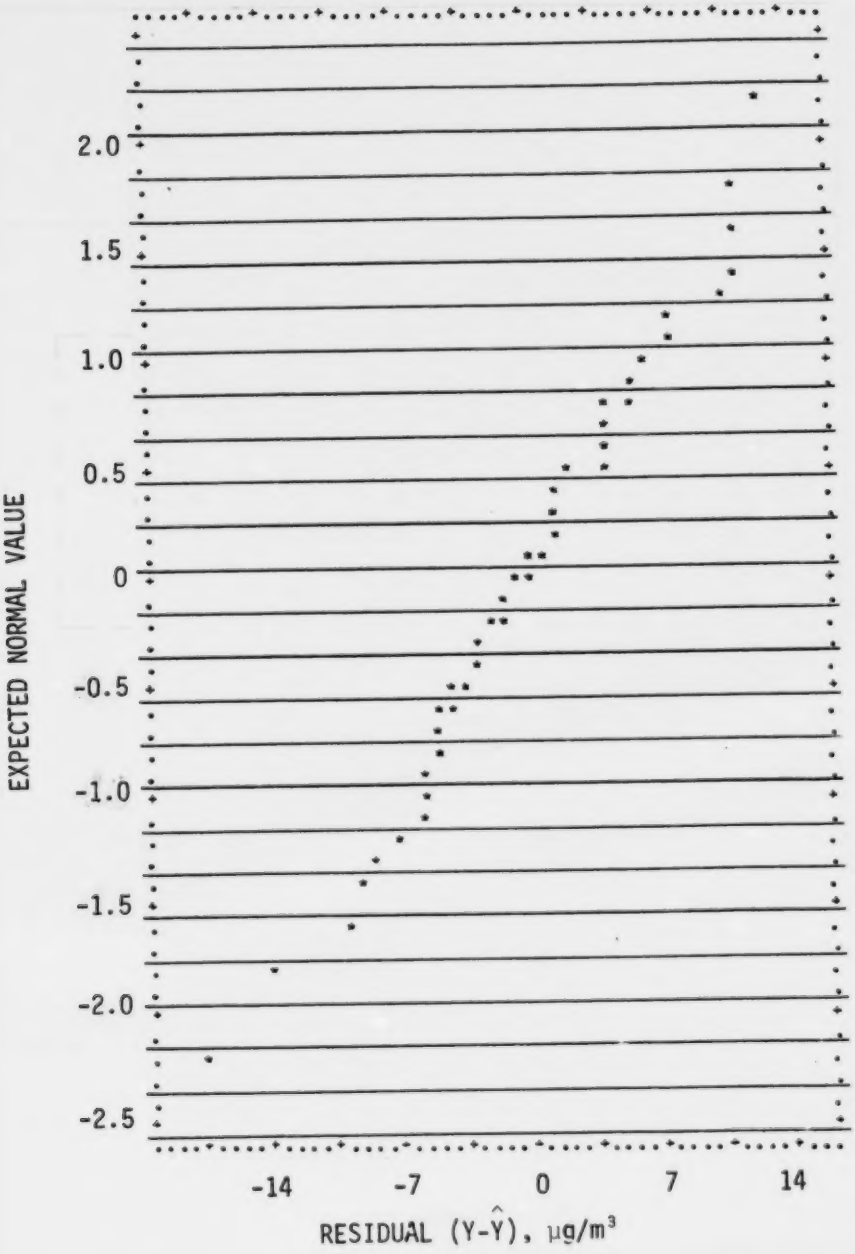


Figure 9. Probability plot of the residuals for Model 2: TOT versus HVL for BIRM data.

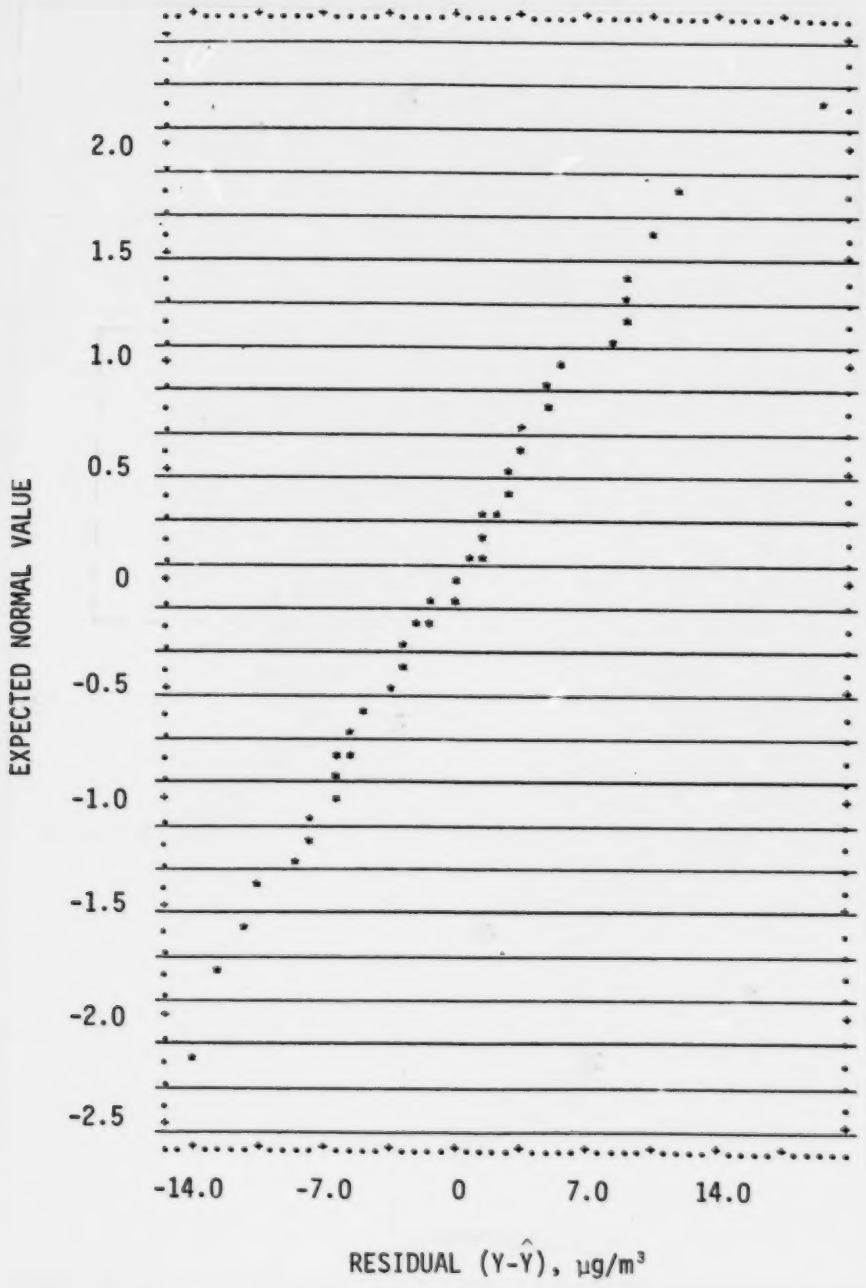


Figure 10. Probability plot of the residuals for Model 2: FNE versus HVL for BIRM data.

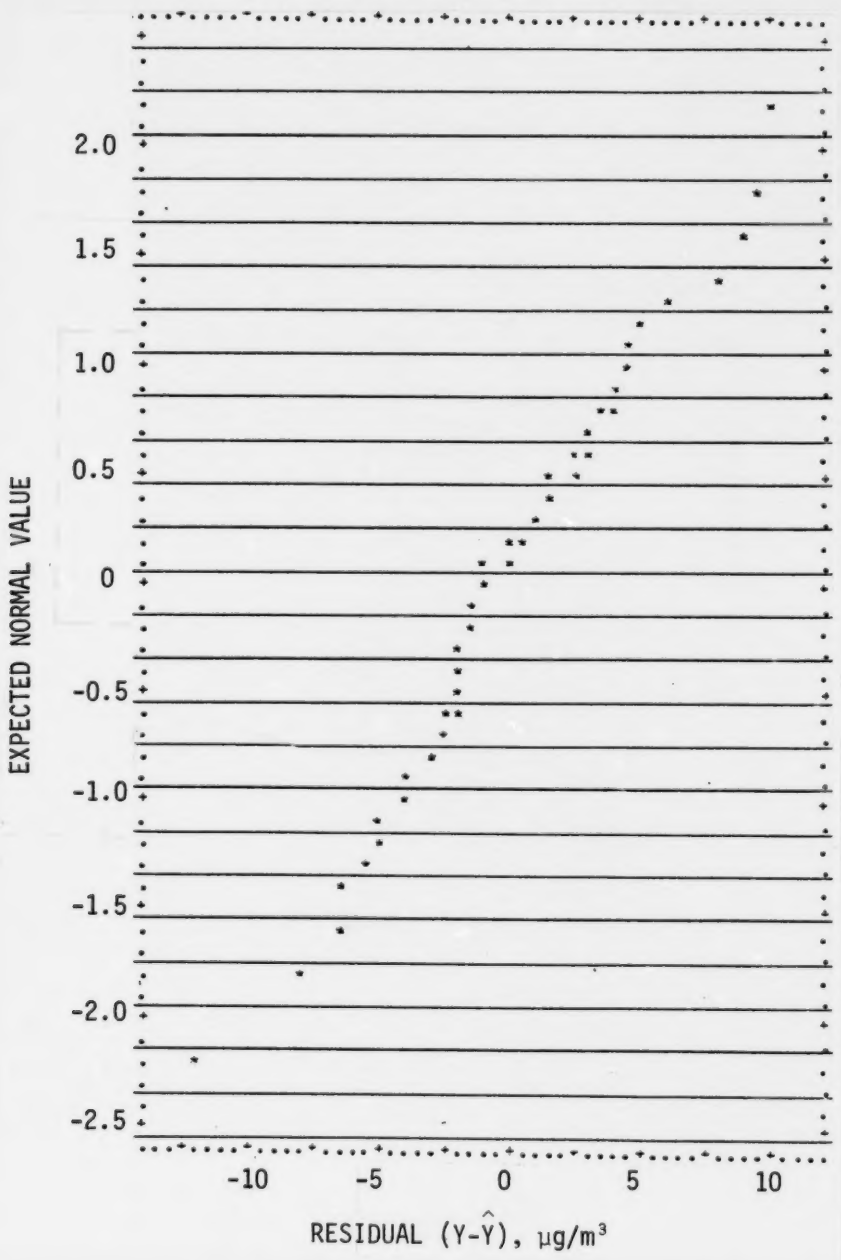


Figure 11. Probability plot of the residuals for Model 2: CRS versus HVL for BIRM data.

approximate normality of the residuals, a necessary condition for the validity of the confidence interval estimation procedures. The R^2 values vary from 0.794 to 0.997 and most exceed 0.95. Although a test procedure proposed by Filliben[11] would be appropriate here, (as suggested by a reviewer), this test was not used. No distinction between the prediction models

$$\hat{Y} = \hat{\beta}X$$

and

$$\hat{Y} = \hat{\beta}_1 X + \hat{\beta}_2 X \log X,$$

based on a comparison of R^2's, is discernible. Why do we consider the two-term model? We do so based on three considerations.

1. In some cases the curvilinear term (X log X) is highly significant.

2. The partial correlations between TOT, FNE and CRS values are computed by the BMDP program, and the relatively large partial correlations (i.e., the correlations after removing the linear effects of HVL) indicate a curvilinear relation between the response variable and the carrier.

3. The linear regression of log Y on log X includes the curvilinear model $Y = \beta_1 X + \beta_2 X \log X$ as a first order approximation when $\hat{\beta}$ is close to unity as derived in Appendix D.

Conclusions with respect to the inclusion of the curvilinear term in the model are tentative. Table 5 presents residual mean square (RESMSQ) and R^2 values for the two models for six sites.

$$R^2 = 1 - \frac{\sum_{i=1}^{n} W_i (Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n} W_i (Y_i - \bar{Y})^2} = \frac{(n-m)\ \text{RESMSQ}}{(n-1)\ \text{Var}(Y)} \qquad (3)$$

TABLE 5. RESIDUAL MEAN SQUARE (RESMSQ) AND R^2 VALUES FOR FOUR MODELS[a]

		Models and weights used in linear regression							
		$Y = \beta X$		$Y = \beta_1 X$		$Y = \beta_1 X + \beta_2 X \log X$		$Y = \beta_2 X \log X$	
		wt = 1		wt = $1/X^2$		wt = $1/X^2$		wt = $1/X^2$	
Site	Y	RESMSQ	R^2	RESMSQ	R^2	RESMSQ	R^2	RESMSQ	R^2
BIRM	SSI	30	0.997	14	0.998	14	0.997		
	FNE	74	0.966	42	0.965	32	0.974		
	TOT	82	0.862	46	0.887	32	0.923		
	CRS	21	0.962	15	0.939	15	0.940		
	SSI	50	0.985	61	0.972	COL		51	0.977
LOSAN1	FNE	136	0.946	101	0.943	COL			
	TOT	142	0.811	86	0.838	COL			
	CRS	30	0.951	25	0.940	COL			
	SSI	52	0.996	56	0.992	COL			
LOSAN2	TOT	563	0.952	101	0.964	102	0.964		
	FNE	479	0.809	98	0.822	101	0.822		
	CRS	172	0.956	66	0.936	67	0.937		
	SSI	no data		no data		COL			
BUFF	TOT	193	0.955	110	0.952	COL			
	FNE	100	0.901	68	0.880	COL			
	CRS	55	0.955	33	0.946	COL			
	SSI	50	0.985	24	0.984	19	0.987		
MIDD	TOT	77	0.976	30	0.977	31	0.978		
	FNE	90	0.853	41	0.936	31	0.910		
	CRS	66	0.944	29	0.963	23	0.944		
	SSI	42	0.958	152	0.959	COL		66	0.982
ELPA	TOT	151	0.968	45	0.913	37	0.931		
	FNE	81	0.727	14	0.701	10	0.794		
	CRS	135	0.953	22	0.920	22	0.924		

[a] R^2 and RESMSQ are given by Equation 3.

where

$$W_i = \frac{1}{X_i^2} .$$

m = number of estimated coefficients in the model (i.e., m=1 for model 1 and m=2 for model 2).

The different weighting schemes should be noted. Although the RESMSQ values in the unweighted model are all larger than those for the weighted models by factors ranging from about 1 to 8, the R^2 values are all approximately equal. This indicates that in the different models the relative change in the sum

$$\sum_{i=1}^{n} W_i(Y_i - \bar{Y})^2$$

is about equal to the relative change in the sum

$$\sum_{i=1}^{n} W_i(Y_i - \hat{Y}_i)^2,$$

Figure 12 is a plot of the fitted SSI versus HVL curve (Model 2) with PHILA2 data. From the plot, it is observed that the average ratio or slope of the fitted curve, SSI/HVL, tends to decrease with increasing HVL values.

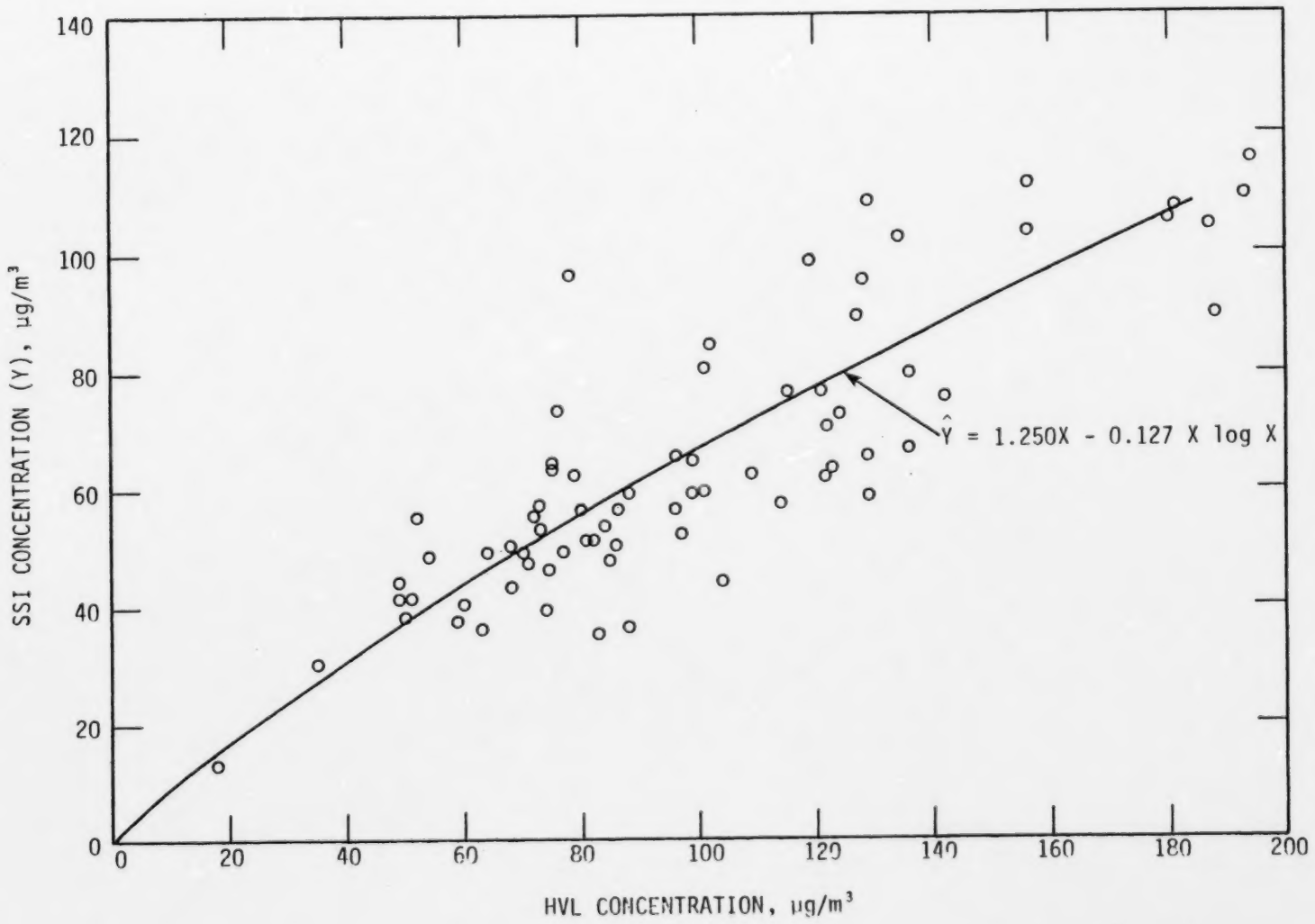


Figure 12. SSI concentration versus HVL concentration using Model 2 with PHILA2 data, (SAROAD No. 397140019).

SECTION 5

STATISTICAL TESTS FOR COMPARING RATIOS AND REGRESSION COEFFICIENTS

In the absence of detailed knowledge of the underlying distributions, one can (i) assume normality (or lognormality) and use the conventional methodology, (ii) set up comparison procedures that are robust, or (iii) use nonparametric statistical techniques. This section discusses conventional and robust methodologies. An examination of nonparametric techniques would exceed the scope of the study.

5.1 CONVENTIONAL METHODOLOGY

If the standard assumptions for regression analysis are appropriate and if IP (= Y) is assumed to be a linear function of TSP or HVL (= X) which passes through the origin (i.e., X = 0, Y = 0), the predicted value (mean) of Y is,

$$\hat{Y} = \hat{\beta}X.$$

Based on the standard regression theory for weighted observations,[12] the average of the ratios is the preferred estimate $\hat{\beta}$ when the standard deviation of Y for given X is proportional to X (Appendix A), that is, $\sigma_{Y|X} = \sigma X$. This appears to be a reasonable assumption based on available data (i.e., the dispersion of the IP values seems to increase proportionally with TSP).[1] Hence, the average ratio $\bar{R}$ will be considered first in the following discussion,

$$\bar{R} = \frac{1}{n} \Sigma\left(\frac{IP}{TSP}\right)$$

and $\bar{R}$ is an estimate of the slope β of the true prediction equation.

When TSP (X) is observed without error and IP (Y) is assumed to be normally distributed for given X with σ (standard deviation of Y given X) proportional to X,

$$\hat{\beta} = \bar{R} = \frac{1}{n} \Sigma\left(\frac{Y}{X}\right).$$

An estimate s_R of σ_R is given by

$$s_R = \left[\frac{\Sigma\,[(\frac{Y}{X}) - \bar{R}]^2}{n - 1}\right]^{1/2} \tag{4}$$

and the standard deviation of $\bar{R}$ is estimated by

$$s_{\bar{R}} = \frac{s}{\sqrt{n}} \tag{5}$$

Under these conditions and the robustness of the t test for relatively large sample sizes

$$t = \frac{\bar{R}-\beta}{s_{\bar{R}}} = \frac{\sqrt{n}\,(\bar{R}-\beta)}{s} \tag{6}$$

has a t-distribution with n - 1 degrees of freedom. Hence, a 95 percent confidence interval for β can be obtained by

$$\bar{R} \pm t_{0.975}\,\frac{s}{\sqrt{n}}\ . \tag{7}$$

In a similar way, two $\bar{R}$'s can be compared. If $\bar{R}_1$ and $\bar{R}_2$ are observed values from two sites or times then

$$t = \frac{\bar{R}_1 - \bar{R}_2}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \tag{8}$$

has a t distribution with $n_1 + n_2 - 2$ degrees of freedom provided $\sigma^2_{R_1} = \sigma^2_{R_2}$, and s is the usual pooled standard deviation, that is,

$$s = \left[\frac{s_1^2(n_1-1) + s_2^2(n_2- 1)}{n_1 + n_2- 2}\right]^{1/2}\ . \tag{9}$$

These test procedures can be illustrated by means of two examples. Tables 6 and 7 show particulate data for 1979 and 1980, respectively) at a single site.[1] If the assumption of equal variances is not valid, and if the calculated t value is close to the critical t (or tabulated t for specified significance level), a more appropriate test procedure may be used as given in the NBS Handbook.[13]

TABLE 6. IP AND TSP DATA FROM PHILA 1 DURING 1979

HVL(X), $\mu g/m^3$	SSI(Y), $\mu g/m^3$	TOT, $\mu g/m^3$	R = SSI(Y)/HVL(X)
95.9	69.1	74.4	0.72
93.7	71.2	66.0	0.76
68.4	55.2	60.0	0.81
54.0	44.2	35.9	0.82
64.5	49.2	55.7	0.76
60.6	48.5	56.8	0.80
43.0	35.4	39.3	0.82
47.8	35.6	40.0	0.74
174.6	118.8	116.5	0.80
50.2	41.2	45.8	0.82
21.6	12.5	21.2	0.58
31.1	14.4	26.6	0.46
57.2	38.5	40.3	0.67
61.2	32.9	44.6	0.54
48.1	36.2	36.0	0.75
27.3	24.9	23.1	0.91
80.5	61.9	57.8	0.77
66.3	57.0	61.9	0.86
36.4	27.8	37.6	0.76
36.1	28.8	32.4	0.80
60.0	49.8	48.4	0.83
72.8	53.6	40.2	0.74
33.8	23.4	42.9	0.69

[a]Source: Reference 1, shown as TSP measurements by HVL sampler and IP measurements by SSI and dichotomous samplers at SAROAD site 39714003, located at 500 South Broad Street and Spruce Street, Philadelphia, Pennsylvania.

TABLE 7. IP AND TSP DATA FROM PHILA 1 DURING 1980

HVL(X)	SSI(Y)	R = SSI/HVL	HVL(X)	SSI(Y)	R = SSI/HVL
59.2	54.4	0.92	59.5	49.9	0.84
71.5	55.2	0.77	59.2	44.7	0.76
43.4	33.1	0.76	119.0	85.0	0.71
64.1	45.8	0.71	60.6	42.6	0.70
62.8	54.2	0.86	50.6	35.9	0.71
61.4	53.6	0.87			
96.2	71.7	0.75			
44.5	34.6	0.78			
47.3	25.2	0.53			
33.9	26.0	0.62			
81.9	56.8	0.69			
52.8	36.0	0.68			
65.5	40.4	0.62			
60.5	44.6	0.74			
70.0	23.6	0.34			
46.4	37.5	0.81			
48.1	31.6	0.66			
30.2	20.9	0.69			
64.9	47.6	0.73			
89.1	69.8	0.78			
57.1	42.3	0.74			
30.7	21.9	0.71			
29.6	17.7	0.60			
44.2	28.7	0.65			
20.3	14.6	0.72			
36.2	22.5	0.62			
44.6	30.9	0.69			
25.2	19.8	0.79			
35.4	19.8	0.56			
71.1	46.2	0.65			
49.4	34.5	0.70			
88.4	65.5	0.74			
57.7	35.3	0.61			
112.5	74.8	0.66			
51.1	42.3	0.83			

Source: Reference 1. Shown are TSP measurements by HVL sampler and IP measurements by SSI sampler at SAROAD site 39714003, located at 500 South Broad Street and Spruce Street, Philadelphia, Pennsylvania.

Example 1. In the first example the 1979 data are used to calculate a 95 percent confidence interval for β_1 (true slope of the regression of SSI(Y) on TSP (X)), given $n_1 = 23$, $\bar{R}_1 = 0.749$ and $s_1 = 0.104$. The 95 percent confidence interval is given by

$$\bar{R}_1 \pm t \frac{s_1}{\sqrt{n_1}}$$

or

$$0.749 \pm (2.074) \frac{0.104}{\sqrt{23}}$$

or approximately,

$$(0.70, 0.80).$$

That is, β_1 (the true average ratio based on 1979 data) should be included in this interval with 95 percent confidence.

Example 2. The second example tests the hypothesis that $\beta_1 = \beta_2$ (i.e., 1979 and 1980 data are equivalent) with data of Tables 6 and 7. (Given for 1980 data that $n_2 = 40$, $\bar{R}_2 = 0.708$, and $s_2 = 0.103$,

$$t = \frac{\bar{R}_1 - \bar{R}_2}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

$$= \frac{0.041}{(0.104)\sqrt{\frac{1}{23} + \frac{1}{40}}} = 1.53 \ (<2.074 \text{ for } \alpha = 0.05 \text{ significance level}).$$

Therefore, the data do not indicate a change in the mean ratio from 1979 to 1980 at the same SAROAD site at the 5 percent level.

The same test procedure can be used to test data from two different sites. In addition an F test can be used to compare the ratios from more than two sites or to obtain simultaneous confidence intervals over all sites.

For small samples (e.g., $n < 20$) it would be preferable to make the log transformation first and then apply the standard statistical tests and confidence interval procedures as described above. The reason for this

approach is based on the assumption that the lognormal distribution generally gives a better fit to the data and the small sample size.

When Model 2 is used to relate IP to TSP for various sites, standard regression analysis techniques should be applied. For example, one can test the equivalence of either one or both of the β_1, β_2 values among sites. The appropriate statistical test is an F test.[10,12]

5.2 ROBUST METHODOLOGY

One means of comparing the ratios for each site would be to use box plots, as described in Section 2. Figures 2 through 5 compared individual values of various ratios. Comparison of the confidence intervals for the medians serves as a robust and conservative procedure for grouping the sites into subsets with essentially equivalent medians within subsets.

SECTION 6

DETERMINATION OF OUTLIERS

Various tests are available for the determination of outliers in a data set containing both IP and TSP (or HVL) measurements. A test for internal consistency (e.g., the Grubbs test)[14] can be applied separately to the IP data and TSP data. This approach, however, does not consider the relationship between the two variables and is not applied when both IP and TSP data are available. Tests deserving more extensive consideration include ones using the IP/TSP ratio, the bivariate normal (lognormal) assumption, and residuals. This section discusses these tests.

6.1 TEST USING THE IP/TSP RATIO

A logical approach to the determination of outliers is to use the IP/TSP ratio as a single variable and to apply an internal consistency check to these values. If both IP and TSP are lognormal, the ratio will be lognormally distributed; otherwise, an approximate form must be used. Two examples follow; assuming first the ratio is normally distributed, and secondly that it is lognormally distributed. The Grubbs test procedure[14] is applied in both examples to data in Table 8 from a previous PEDCo study[5] as briefly described in Section 2.

Example 1. The following statistics are computed from the data in Table 8:

$n = 37$

$\bar{R} = 0.71$ (average ratio, IP/TSP)

$s = 0.21$ (standard deviation of the ratio)

The smallest value of the IP/TSP (Y/X) ratio is $R_1 = 0.14$ and is a potential data anomaly or outlier. Applying the Grubbs test to this value yields

TABLE 8. INHALABLE PARTICULATE (IP) AND TOTAL SUSPENDED PARTICULATE (TSP) CONCENTRATIONS AND THEIR RATIO[a]

Date	IP concentration (Y) $\mu g/m^3$	TSP concentration (X) $\mu g/m^3$	IP/TSP (Y/X) ratio
6-13	84	84	1.00
6-14	72	96	0.75
6-15	35	47	0.74
6-16	28	40	0.70
6-17	35	48	0.73
6-19	65	133	0.49
6-20	30	49	0.61
6-21	42	58	0.72
6-23	56	93	0.60
6-24	58	58	1.00
6-26	79	87	0.91
6-28	45	83	0.54
6-29	48	71	0.68
6-30	59	120	0.49
7-03	122	244	0.50
7-07	78	205	0.38
7-08	77	99	0.78
7-09	61	46	1.33
7-11	60	73	0.82
7-12	66	80	0.82
7-14	69	87	0.79
7-15	90	130	0.69
7-16	67	131	0.51
7-17	93	169	0.55
7-18	51	84	0.61
7-19	76	90	0.84
7-20	84	105	0.80
7-21	34	59	0.58
7-22	97	170	0.57
7-23	51	56	0.91
7-24	41	59	0.69
7-25	47	52	0.90
7-26	76	86	0.88
7-28	145	242	0.60
7-29	29	207	0.14
7-30	49	65	0.75
7-31	76	94	0.81
Average	64.2	100.3	0.71
Std dev	25.5	54.2	0.21
n	37	37	37

[a]Source: Reference 5.

$$T = \frac{\bar{R} - R_1}{s} = \frac{0.71 - 0.14}{0.21} = 2.77. \quad (10)$$

The critical value for $\alpha = 0.05$ risk is 2.835 and for $\alpha = 0.10$ risk is 2.65. Thus, 0.14 is a borderline value and would not be considered questionable on the basis of this statistical test. The following lognormal application is also used because previous results suggest that in general this approach is preferred.

Example 2. The following statistics are computed for the logarithms of the ratio data in Table 8.

$n = 37$

$\bar{R}' = -0.397$ [average of the ln (Y/X)]

$s' = 0.363$ [standard deviation of the ln (Y/X)]

The value in question is 0.14, the ln 0.14 = -1.97. Hence,

$$T = \frac{-0.397 + 1.97}{0.363} = 4.33.$$

This value exceeds $T_{0.001} = 3.631$ tabulated by Grubbs[14] and thus indicates that the 0.14 value is very questionable. This is an unusual example where the value is indicated as questionable after the log transformation as compared to the calculation using untransformed data. It appears that this usually will be the case for very small values of the ratio. Hence, knowing the distribution of the IP/TSP ratio or the individual IP and TSP measurements is important.

This ratio test does not indicate which of the two measurements (IP = 29 or TSP = 207 $\mu g/m^3$) is subject to question but only that the IP/TSP ratio (0.14) is suspect. If one applies separate tests to IP and to TSP, one finds that neither the IP nor the TSP in suspect. Actually, there are larger TSP values than 207 $\mu g/m^3$ and several IP values close to 29 $\mu g/m^3$ (e.g., 30, 34, 35, 35).

6.2 TEST USING THE BIVARIATE NORMAL (LOGNORMAL) ASSUMPTION

Use of the IP/TSP ratio implies that IP is linearly related to TSP and that a fitted line should pass through the origin. Other approaches are applicable to variables subject to a linear relationship whether or not the fitted line passes through the origin. One procedure is described in the data validation report[15] under the heading "Intersite Correlation Test," and can be used for comparing two correlated parameters or measurements. This procedure may be applied to the raw data or to the log transform of the data. Both cases will be described here. A 95 percent probability ellipse will be obtained in each case to aid in identifying questionable values. A second procedure based on as examination of the residuals will be described in Section 6.3.[16]

Case 1. The data for both IP (Y) and TSP (X) of Table 8 are used in the analysis for an outlier. Results of the analysis (excluding the potential outlier) are as follows:

$$n = 36$$

$$\Sigma X = 3503 \qquad \Sigma Y = 2346$$

$$\Sigma X^2 = 435{,}073 \qquad \Sigma Y^2 = 174{,}954$$

$$\Sigma XY = 267{,}400$$

$$\bar{X} = 97.3 \qquad \bar{Y} = 65.2$$

$$s_X = 51.9 \qquad s_Y = 25.1$$

$$r^2 = 0.736$$

Assuming IP can be predicted by a linear relationship (not necessarily through the origin),

$$\hat{Y} = a + bX$$

then

$$a = 24.76 \qquad b = 0.415 \qquad s_{Y\,X} = 13.1$$

$$s_a = 4.7 \qquad s_b = 0.043$$

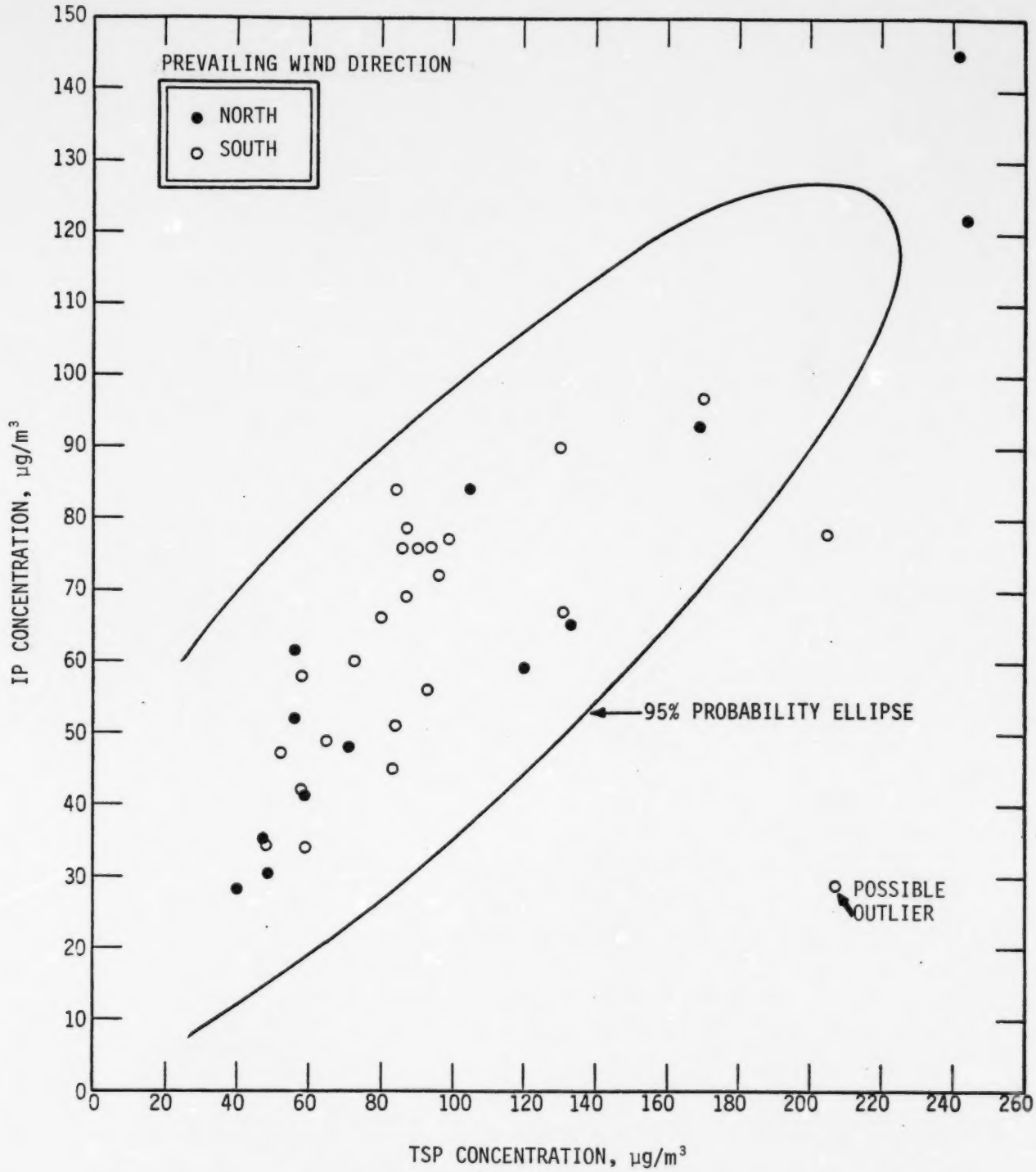


Figure 13. IP versus TSP for collocated SSI and HVL samplers.

The data are plotted in Figure 13; the probability ellipse is obtained by the procedure described in the data validation report.[15] The equation of the ellipse is

$$\frac{1}{1-r^2}\left[\left(\frac{X-\bar{X}}{s_X}\right)^2 - 2r\left(\frac{X-\bar{X}}{s_X}\right)\left(\frac{Y-\bar{Y}}{s_Y}\right) + \left(\frac{Y-\bar{Y}}{s_Y}\right)^2\right] = \chi^2_{2DF,\ 0.95} = 5.99 \tag{11}$$

Substituting the above results into Equation (11) yields

$$\frac{1}{0.264}\left[\left(\frac{X-97.3}{51.9}\right)^2 - 2(0.858)\left(\frac{X-97.3}{51.9}\right)\left(\frac{Y-65.2}{25.1}\right) + \left(\frac{Y-65.2}{25.1}\right)^2\right] = 5.99$$

The ellipse falls (is inscribed) within a rectangle with $(\bar{X}, \bar{Y})$ as its center and with sides of the following lengths along the X and Y axes:

$$2s_X\chi = 2(51.9)\sqrt{5.99} = 254.0 \text{ units}$$

$$2s_Y\chi = 2(25.1)\sqrt{5.99} = 122.9 \text{ units.} \tag{12}$$

Because the ellipse extends beyond the zero boundary, the log transform of the data should be considered. Based on the results one would certainly question the point IP = 29 $\mu g/m^3$, TSP = 207 $\mu g/m^3$. The other data points outside the 95 percent probability ellipse would probably fall within a 99 percent probability ellipse.

Case 2. In the second case the (natural) logarithms of the data in Table 8 are used in the analysis for outliers. The results of the data analysis are as follows, where the superscript prime (′) denotes the log of the variable

$$n = 36$$

$$\Sigma X' = 160.741 \qquad \Sigma Y' = 147.865$$

$$\Sigma(X')^2 = 725.249 \qquad \Sigma(Y')^2 = 612.409$$

$$\Sigma X'Y' = 665.524$$

$$\bar{X}' = 4.344 \qquad \bar{Y}' = 3.996$$

$$s_{X'} = 0.464 \qquad s_{Y'} = 0.381$$

$$(r')^2 = 0.736$$

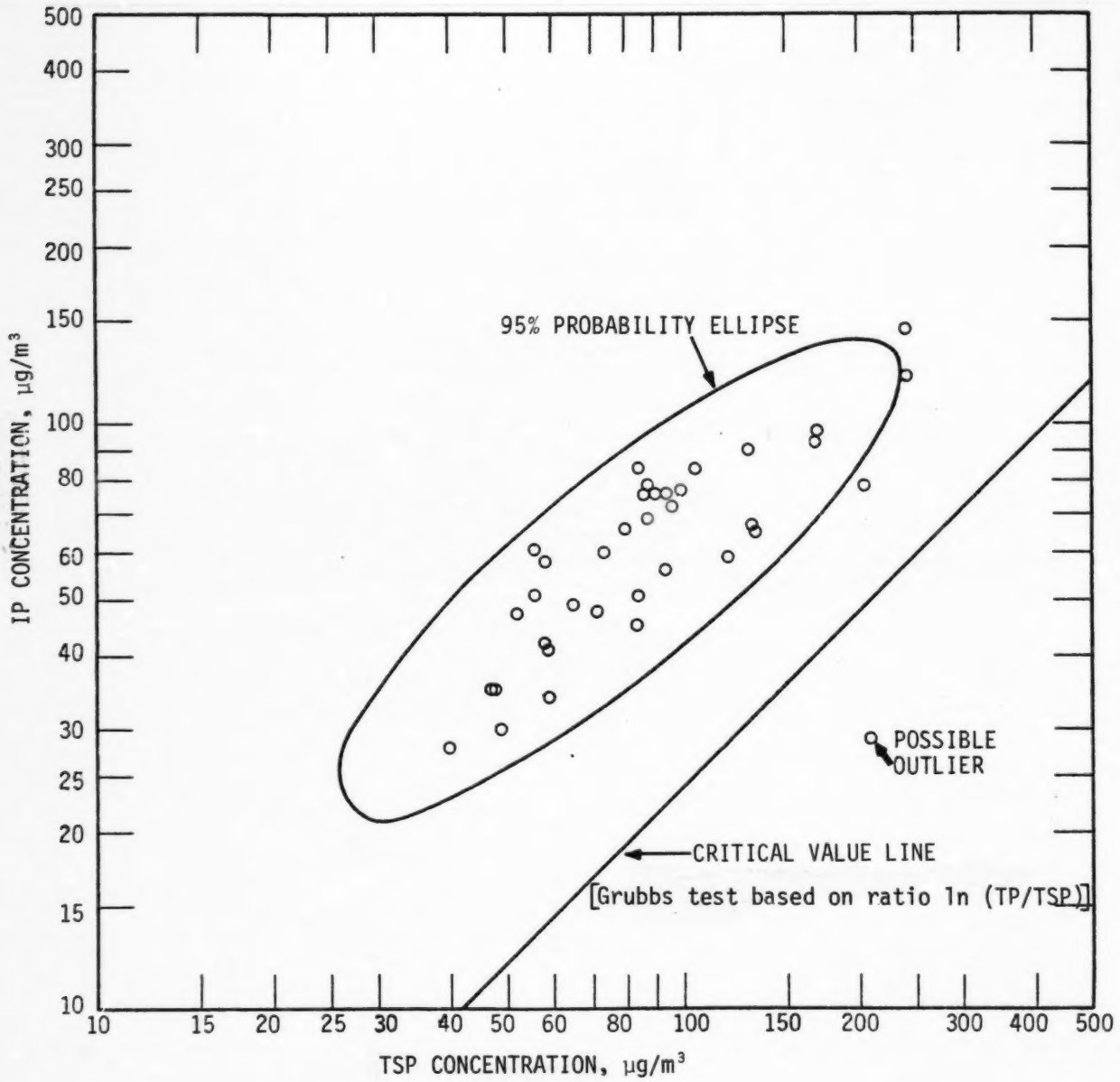


Figure 14. Log IP versus log TSP for collocated SSI and HVL samplers.

The two correlation coefficients, r and r', for the original and transformed data, are equal by coincidence. Assuming that the ln IP and ln TSP have a bivariate normal distribution, then the following approximation holds

$$\frac{1}{0.264}\left[\left(\frac{X'-4.344}{0.464}\right)^2-2(0.858)\left(\frac{X'-4.344}{0.464}\right)\left(\frac{Y'-3.996}{0.381}\right)+\left(\frac{Y'-3.996}{0.381}\right)^2\right] = 5.99.$$

The 95 percent probability ellipse defined by the above equation is given in Figure 14. In addition the critical value line for the ratio ln (IP/TSP) is given for $\alpha = 0.05$. The difference between the probability ellipse and the ratio test (indicated by the straight line) is explained in part by the fact that the ellipse assumes that the means, variances, and the correlation are known. Another procedure can be applied which bypasses this assumption, using the multivariate T (or T^2) distribution (the multivariate analog of the t distribution).[17]

6.3 TEST USING RESIDUALS (REGRESSION APPROACH)

Another procedure to detecting outliers is based on the results of Hoaglin and Welsch.[16] The following strategy is to be used.

1. First, examine the diagonal elements of (h_{ii} or h_i) of the "hat" matrix(H)[16] for data points that significantly affect the results (referred to as high-leverage points).

2. Then examine the residuals (actually the Studentized residuals values r_i^*)[16] for potential outliers.

The details of the computation are given in an example for the data in Table 8. For the two variable case, the h_i are defined as

$$h_i = h_{ii} = \frac{1}{n} + \frac{X_i - \bar{X}^2}{\sum_{h=1}^{n} (X_h - \bar{X})^2}, \qquad (13)$$

and the r_i^* (Studentized residual value) as

$$r_i^* = \frac{r_i}{s_{(i)}\,(1-h_i)^{1/2}},$$

where

r_i = the residual of the ith value Y_i from the predicted value $\hat{Y}_i$ using all of the data except the point (X_i, Y_i), which is a potential outlier. Thus, there are n-1 data points.

$s_{(i)}$ = the standard deviation of the residuals based on the use of all of the data excluding point i.

Data from Table 8 excluding the point (29,207) were used to determine a best fit prediction equation in the logarithms of the variables, that is,

$$\ln IP = 0.965 + 0.704 \ln TSP$$

$$r^2 = 0.736$$

$$s_{Y\,X} = 0.20 \text{ (standard deviation of the residuals)}$$

$$n = 36$$

$$\bar{X}' = 4.465, \quad \bar{Y}' = 4.107$$

Using these data, h_i and r_i^* are calculated as follows:

$$h_i = \frac{1}{36} + \frac{(0.868)^2}{7.534} = 0.128$$

$$r_i = 3.367 - 4.719 = -1.352$$

$$r_i^* = \frac{-1.352}{0.20\ (1-0.128)^{1/2}} = 7.24.$$

The statistic r_i^* has a t-distribution with n-3 = 33 degrees of freedom. Hence, we would infer the residual to be significantly different from zero.

A rule of thumb for significance of the h_i is that it exceeds 4/n for the two variable case. Because the observed value 0.128 > 4/36 = 0.11, the data point (29,207) is a potential outlier. A statistical test for h_i is available when the regression model includes a constant term. For any single h_i, the statistic

$$F_{p-1,\ n-p} = \frac{(n-p)(h_i-1/n)}{(p-1)(1-h_i)}$$

has an F distribution with p-1 and n-p degrees of freedom. In the present example, n=36, p=2, h_i = 0.128, hence

$$F_{1,35} = \frac{34\ (0.128 - \frac{1}{36})}{1\ (0.128)} = 3.91.$$

Because the tabulated value of F for (1,35) degrees of freedom is 4.12, the value is borderline and does not necessarily indicate the point to be an outlier, whereas the test for r_i^* strongly suggests an outlier. Both the r_i^* and $F_{p-1,\ n-p}$ tests should be applied because they augment one another in the identification of an outlier.[16]

6.4 CONCLUSION

The test using the IP/TSP ratio is the simplest test and would be recommended for this reason. In addition, the t test is a robust test (i.e., not very sensitive to deviations from assumptions). Use of the logarithmic transformation, however, would be preferred because the log-normality tests of Section 3 generally suggest that the lognormal is a better fit than the normal distribution. The scatter plot is also recommended for a graphical check when the statistical test or tests (if more than one technique is used) suggest a possible outlier. Other useful graphs are a plot of the residuals versus TSP(Y) values, a plot of residuals $r_i = Y_i - \hat{Y}_i$ versus $\hat{Y}_i$, and a plot of residuals on normal probability paper.[16]

SECTION 7

PROCEDURE FOR SELECTION OF IP MONITORING SITES BASED ON TSP DATA

One problem of practical interest is the selection of IP monitoring sites. It is assumed for purposes of discussion that the IP standard is known, (e.g., the annual average IP concentration must be less than 60 $\mu g/m^3$ and all daily averages must be less than 200 $\mu g/m^3$). This standard is a hypothetical one. Based on available TSP data at a specific site and other information concerning IP/TSP data at "similar" sites, one is to select values for annual average TSP concentration and daily average TSP concentration that should not be exceeded based on a level of risk α. At sites where TSP (or the appropriate confidence limits) do not exceed the specified values, an IP monitor is not required. At other sites where one or both of the specified TSP values is exceeded, an IP monitor is presumed to be necessary. This section contains a brief discussion of an approach and an example to illustrate the procedure.

Suppose that IP(Y) is a linear function of TSP(X), passing through the origin (X = 0, Y = 0), that is,

$$\hat{Y} = \hat{\beta}X = \bar{R}X$$

based on results from Section 5. Also, assume that the standard deviation of the residuals is proportional to X, and that there is no error in the observed X. Let $\hat{Y} = \bar{R}X$ be the prediction equation for the mean of IP given TSP applying the standard regression techniques. From the results in Section 5, $s^2_{\bar{R}}$ (the estimated variance of $\bar{R}$) is given by

$$s^2_{\bar{R}} = \frac{s^2}{n} \qquad (14)$$

where s^2 is the variance of R,

$$s^2 = \frac{\Sigma[(\frac{Y}{X}) - \bar{R}]^2}{n-1} . \quad (15)$$

It is assumed that no IP (Y) data are available and hence s^2 cannot be obtained at the site under investigation. Thus, s^2 must be estimated from data from similar sites where both IP and TSP are available. Assuming X is known without error, the variance of $\hat{Y}$ given X is

$$\sigma^2_{\hat{Y}} = X^2\sigma^2_{\bar{R}} , \quad (16)$$

Hence, a confidence interval for the mean IP concentration for given X can be obtained by

$$\hat{Y} \pm t_{0.975}\, s_{\hat{Y}} = \hat{Y} \pm t_{0.025}\, s_{\hat{Y}} \quad (17)$$

where

$$s_{\hat{Y}} = X\, s_{\bar{R}}. \quad (18)$$

For example, using the 1979 data from Table 6, ($n = 23$, $\bar{R} = 0.75$, $s = 0.104$), the following can be determined:

$$\hat{Y} = 0.75X$$

$$s_{\hat{Y}} = \frac{0.104}{\sqrt{23}}\, X.$$

A 95 percent confidence interval for the mean IP is

$$\hat{Y} \pm t_{0.975}\, s_{\hat{Y}}$$

or

$$0.75X \pm 2.074\, (\frac{0.104}{\sqrt{23}})\, X$$

or

$$0.75X \pm 0.045X = X(0.75 \pm 0.45).$$

Hence, if the observed mean X is 100, then the 95 percent confidence interval for the mean IP is 70.5 to 79.5 $\mu g/m^3$. Without any question the decision would be to install an IP monitor. If the observed mean TSP concentration were 70 $\mu g/m^3$, then the 95 percent confidence interval would be 52.5 ± 3.2 $\mu g/m^3$, or 49.3 to 55.5 $\mu g/m^3$. Since 55.5 $\mu g/m^3$ is less than the assumed standard of 60 $\mu g/m^3$, no monitor would be required based on the annual mean level. We still need to investigate the individual daily averages, however, for possible exceedance of the hypothetical IP daily standard of 200 $\mu g/m^3$.

An individual daily IP value for a given X is given by

$$Y = \bar{R}X + e$$

where e is the error of measurement with variance proportional to X^2 (i.e., $\sigma^2(e) = \sigma^2X^2$), and thus

$$\sigma_Y^2 = \sigma_{\bar{R}}^2X^2 + \sigma^2X^2. \tag{19}$$

From the results of Section 5,

$$\sigma_{\bar{R}}^2 = \frac{\sigma_R^2}{n} = \frac{\sigma^2}{n}, \text{ and thus}$$

$$\sigma_Y^2 = \sigma^2X^2[\frac{1}{n} + 1].$$

Thus

$$s_Y = sX\,[\frac{1}{n} + 1]^{1/2}$$

and a 95 percent confidence interval for a given X is

$$Y \pm t_{0.975}\, sX[1 + \frac{1}{n}]^{1/2}. \tag{20}$$

If all TSP values (X's) are less than 200 $\mu g/m^3$ for the site under investigation, then for X = 200 $\mu g/m^3$, the 95 percent confidence interval for Y is 150 ± 44.1 $\mu g/m^3$, or 106 to 194 $\mu g\ IP/m^3$. This is a conservative approach in that the interval is based on the largest X. Because 194 $\mu g/m^3$ is less than 200 $\mu g/m^3$ it can be inferred that no monitor would be necessary (i.e., we are accepting a very small risk by not locating an IP monitor at this site because the chance is very small that we would observe a value exceeding the hypothetical standard.

Another problem of interest is that of obtaining a confidence interval for the average IP/TSP ratio over a population of sites. Suppose that the ratio estimate $\bar{R}$ for the sth site is $\bar{R}_s$ (s = 1, 2, ..., S). Assume that these sites have been randomly sampled from a specific population of sites (e.g., all sites in one EPA Region). Assume that $\bar{R}_s$ (s = 1, ..., S) are a random sample of observations from a normal population with mean μ and variance $\sigma^2_{\bar{R}}$. A grand mean ratio ($\bar{\bar{R}}$) can be computed as follows:

$$\bar{\bar{R}} = \frac{1}{N} \sum_{s=1}^{S} n_s \bar{R}_s \ , \tag{21}$$

where

$$N = \sum_{s=1}^{S} n_s. \tag{22}$$

An estimate of the variance of $\bar{\bar{R}}$ is given by

$$\widehat{Var}\ (\bar{\bar{R}}) = \frac{1}{S-1} \sum_{s=1}^{S} n_s\ (\bar{R}_s - \bar{\bar{R}})^2. \tag{23}$$

An approximate 95 percent confidence interval for μ is given by

$$\bar{\bar{R}} \pm Z_{1-\alpha/2} \sqrt{\widehat{Var}(\bar{\bar{R}})}. \tag{24}$$

The quantity $Z_{1-\alpha/2}$ is the 100 $(1-\alpha/2)$ percentile of the standard normal distribution. Given TOT (dichotomous sampler) data from the 11 locations sampled, the grand mean ratio is $\bar{\bar{R}}$ = 0.70, and the 95 percent confidence interval is 0.59 to 0.80.

SECTION 8

CONCLUSIONS

Based on the study of eleven data sets described in Section 2 and the data set from another PEDCo study,[1,5] the following conclusions are given.

1. Both the IP and TSP data are better approximated by the lognormal distribution than by the normal distribution. This has been an accepted conclusion for TSP data over many years; however, the IP data also seem to be better approximated by the lognormal distribution. We are <u>not</u> inferring that IP data are lognormally distributed, because a comparison with other distribution forms was not made.

2. If both the IP and TSP data are approximately lognormally distributed, the log of the IP/TSP ratio is approximately normally distributed [i.e., log (IP/ TSP) = log IP - log TSP, the difference of two normal variables]. Thus, many calculations with the data are simplified as noted in subsequent conclusions.

3. Box plots of the IP/TSP ratio for the 11 basic data sites vary considerably among sites and usually the sites can be divided into two or three subsets with nearly equal medians (actually not significantly different medians) within a subset but significantly different medians between subsets. The box plots are a convenient means of summarizing the data from many sites.

4. Two standard tests were performed to test normality (or lognormality) and the results were nearly identical and hence, the values for only one test (Cramér-von Mises W^2) are given.

5. The standard regression analysis requires that the "independent" variable(s) be known without error. The effect of error in TSP measurements on the determination of the IP-TSP relationships was investigated. This effect was found to be small.

6. Two basic models are proposed for relating IP to TSP, the corresponding prediction equations are:

$$\hat{Y} = \bar{R}X$$

where

$$\bar{R} = \frac{1}{n} \Sigma \frac{IP}{TSP}, \text{ the average of the ratios,}$$

and

$$\hat{Y} = \hat{\beta}_1 X + \hat{\beta}_2 X \log X.$$

Model 2 (the two-parameter model) does improve the fit relative to Model 1 in many cases and Model 2 is a first order approximation to

$$\log Y = \hat{\alpha} + \hat{\beta} \log X$$

provided β is nearly equal to unity ($\beta \tilde{} 1$) as shown in Appendix D. The two parameter model is preferred when a very good prediction relationship is desired. For many applications, however, the simpler prediction equation $\hat{Y} = \bar{R}X$ is adequate; furthermore, many standard statistical procedures can be used to compare data from different sites.

7. Weighted regression should be used to improve the fit of either model (i.e., use weights equal to $1/TSP^2$).

8. Conventional statistical tests for comparing ratios and obtaining confidence intervals can be used. Results can be improved by using logarithms of ratios in all calculations and assuming that the log ratios are normally distributed. Examples are given to illustrate these test procedures.

9. The recommended outlier test is the log ratio with the standard Grubbs test, and a plot of the data for significant results from this test. Other tests which would be preferable if a computerized procedure were used (e.g., a probability ellipse, assuming log IP and log TSP have a bivariate normal distribution). In addition, some standard regression techniques are given for the determination of outliers in case a more complex relationship (e.g., Model 2) is used.

10. A procedure is given for obtaining a confidence interval for the grand mean ratio $\bar{R}$ from a sampled population of sites. In addition, a procedure is given for selecting sites at which an IP monitor should be placed on the basis of TSP data and an IP standard.

11. Two or more measurement techniques can be compared for their relative precisions as indicated in Appendix C.

SECTION 9

REFERENCES

1. Watson, J. G., J. C. Chow and J. J. Shaw. Analyses of Measurements From the Inhalable Particulate Matter Sampling Network, Unpublished Report on Contract EPA 68-02-2542, Task No. 6, Environmental Research and Technology, Inc. A394-140, July 1981.

2. Pace, T. G. Preliminary Characterization of IP/TSP Ratio. U.S. Environmental Protection Agency, Research Triangle Park, North Carolina, May 1979.

3. Tukey, J. W. Exploratory Data Analysis. Addison-Wesley Publishing Co., Reading, Massachusetts, 1977.

4. McGill, et al. Variation of Box Plots. The American Statistician, 32:12-17, 1978.

5. Murphy, W. Impact of Stone Quarry Operations on Particulate Levels. Prepared under Contract No. 10.085, PEDCo, Kansas City, Missouri, September 1980.

6. Stephens, M. A. EDF Statistics for Goodness of Fit and Some Comparisons. Journal of the American Statistical Association, 60:730-737, September 1974.

7. Biomedical Computer Programs P-Series. University of California Press, Berkeley, 1979.

8. Madansky, A. The Fitting of Straight Lines when Both Variables Are Subject to Error. Journal of the American Statistical Association, 173-205, March 1959.

9. Kunitomo, N. Asymptotic Expansions of the Distributions of Estimators in a Linear Functional Relationship and Simultaneous Equations. Journal of the American Statistical Association, 173-205, March 1959.

10. Chatterjee, S., and B. Price. Regression Analysis by Example. John Wiley and Sons, Inc., New York, 1977.

11. Filliben, J. J. The Probability Plot Correlation Coefficient Test for Normality. Technometrics, 17:111-117, 1975.

12. Hald, A. Statistical Theory With Engineering Applications, John Wiley and Sons, Inc. New York, New York, 1952.

13. Natrella, M. G., Experimental Statistics, Ordnance Corps Pamphlet, ORDP 20-110, June 1962.

14. Grubbs, F. E. Errors of Measurement, Precision, Accuracy and the Statistical Comparison of Measuring Instruments. Technometrics, 15(1):53-66, February 1973.

15. Nelson, A. C., Jr., D. W. Armentrout, and T. R. Johnson. Validation of Air Monitoring Data. EPA-600/4-80-030, June 1980.

16. Hoaglin, D. C., and R. E. Welsch. The Hat Matrix in Regression and ANOVA. The American Statistician, 32(1), February 1978.

17. Anderson, T. W., An Introduction to Multivariates Analysis, John Wiley & Sons, Inc., NY, 1958.

18. Maloney, C. J., and S. C. Rastogi. Significance Test for Grubbs' Estimators. Biometrics, 26:671-676, 1970.

19. Morgan, W. A. A Test for the Significance of the Difference Between Two Variances in a Sample From a Normal Bivariate Population, Biometrika, 31:13-19, 1939.

20. Land, C. E. Confidence Interval Estimation Methods. Technometrics, 14:145-158, 1971.

APPENDIX A

WEIGHTED REGRESSION ANALYSIS

Assume that Y is a linear function of X; that is,

$$Y = \beta X, \tag{A-1}$$

where X is observed without error and

$$\sigma_{Y|X} = \sigma f(X). \tag{A-2}$$

If a weighted regression analysis is used,[11] then

$$\bar{X} = \frac{\Sigma w_i X_i}{\Sigma w_i} = \text{weighted average of X values} \tag{A-3}$$

$$b = \frac{\Sigma w_i X_i Y_i}{\Sigma w_i X_i^2} = \text{estimate of } \beta \tag{A-4}$$

$$w_i = \frac{1}{f^2(X_i)} = \text{weight} \tag{A-5}$$

$$Y_p = bX = \text{predicted mean value for Y} \tag{A-6}$$

Case 1. If $f(X) = X$, then $w_i = \frac{1}{X_i^2}$, and

$$b = \frac{1}{n} \Sigma\left(\frac{Y_i}{X_i}\right). \tag{A-7}$$

Case 2. If $f(X) = C$ (where C is a constant)

$$b = \frac{\Sigma X_i Y_i}{\Sigma X_i^2}. \tag{A-8}$$

Case 3. If $f(X) = \sqrt{X}$, then $w_i = \frac{1}{X_i}$, and

$$b = \frac{\Sigma Y_i}{\Sigma X_i} = \frac{\bar{Y}}{\bar{X}}. \tag{A-9}$$

APPENDIX B

ERRORS IN BOTH VARIABLES

Two papers on the estimation of parameter(s) of a linear functional relationship (errors in both variables) serve as a basis for this discussion.[8,9] One is a summary paper of work done prior to 1955,[8] and the other is a recent paper on approximate distribution of estimators.[9] Many other papers on functional relationship, errors in both variables, and related topics are referenced in these two papers. This appendix applies approaches from the papers to the problem of interest in this report. In this appendix, lowercase x's and y's denote the observations with errors, whereas uppercase X's and Y's denote the corresponding true values. Although this notation differs from that in the text of the report, no confusion should result because this appendix is self-contained.

Approach 1. One approach to the problem of errors in both variables is the following:

Assumptions: $x_i = X_i + u_i$

$y_i = Y_i + v_i$

$Y_i = \alpha + \beta X_i$

u_i is $N(o,\sigma_u^2)$ (i.e., normally distributed with mean o and variance σ_u^2)

v_i is $N(o,\sigma_v^2)$

u_i, v_i are the errors of observation of x_i and y_i.

If σ_v^2 is known,

$$\hat{\beta} = \frac{S_{yy} - n\sigma_v^2}{S_{xy}} . \qquad \text{(B-1)}$$

If σ_u^2 is known,

$$\hat{\beta} = \frac{S_{xy}}{S_{XX} - n\sigma_u^2} \qquad \text{(B-2)}$$

If $\lambda = \sigma_v^2/\sigma_u^2$ is known,

$$\hat{\beta} = \frac{S_{yy} - \lambda S_{xx} + \{[S_{yy} - \lambda S_{xx}]^2 + 4\lambda[S_{xy}]\}^{1/2}}{2 S_{xy}} \tag{B-3}$$

where

$$\begin{aligned} S_{yy} &= \Sigma(y_i - \bar{y})^2 \\ S_{xx} &= \Sigma(x_i - \bar{x})^2 \\ S_{xy} &= \Sigma(x_i - \bar{x})(x_i - \bar{y}) \end{aligned} \tag{B-4}$$

Then

$$\hat{\alpha} = \bar{y} - \hat{\beta}\,\bar{x}. \tag{B-5}$$

Given the data from Table 6 ($n = 23$, $S_{yy} = 9023.70$, $S_{xx} = 11{,}256.35$, and $S_{xy} = 9557.45$), standard regression analysis yields the following results: $\hat{\alpha} = 9.95$, $\hat{\beta} = 0.8491$, $r^2 = 0.8993$, and $s_{y|x} = 6.58$.

Given the above results for errors in both variables, $\hat{\beta} = 0.884$ if $\sigma_v^2 = 25$, $\sigma_u^2 = 25$; $\hat{\beta} = 0.845$ if $\sigma_u^2 = 25$; and $\hat{\beta} = 0.890$ if $\lambda = 1$. These values are to be compared with the least-squares $\hat{\beta}$ value of 0.849.

Asymptotic distributions of the least-squares estimator ($\hat{\beta}_L$) and the maximum-likelihood estimator ($\hat{\beta}_M$) have been obtained.[9] On the whole, the cumulative distribution function (cdf) of $\hat{\beta}_L$ is biased downward and is more concentrated around $[c/(1 + c)]\beta$ than the cdf of $\hat{\beta}_M$ around β for a relatively large sample size (n). The constant c is the ratio of the variance of the x_i around their mean to the σ_u^2. For example, a c value of 9 is reasonable, and hence $\hat{\beta}_L$ is concentrated around 0.9 β, that is, it is biased downward by 0.1 β.

Approach 2. Another approach is to obtain the mean and variance of the estimator (b) obtained from weighted regression theory, that is,

$$b = \frac{\Sigma(y_i/x_i)}{n} \tag{B-6}$$

where

$$x_i = X_i + u_i,\ \sigma_{u_i}^2 = \sigma_u^2 X_i^2$$

$$y_i = Y_i + v_i, \quad \sigma^2_{v_i} = \sigma^2_v X_i^2$$

$$Y = \beta X.$$

The estimator b can be expressed as

$$b = \frac{1}{n} \Sigma (\beta + \frac{v_i}{X_i})(1 - \frac{u_i}{X_i} + \frac{u_i^2}{X_i^2} - \cdots). \tag{B-7}$$

Expected value of b [E(b)]

$$E(b) \cong \beta + \frac{\beta}{n} \Sigma \frac{E(u_i^2)}{X_i^2} \tag{B-8}$$

If $\sigma^2(u_i) = E(u_i^2) = \sigma^2_u$, then

$$E(b) \cong \beta(1 + \frac{\sigma_u^2}{n} \Sigma \frac{1}{X_i^2}) \tag{B-9}$$

If $\sigma^2(u_i) = \sigma^2_u X_i^2$, then

$$E(b) \cong \beta + \beta \sigma^2_u = \beta (1 + \sigma^2_u) \tag{B-10}$$

For example, σ^2_u might be $(0.05)^2$, indicating that b is biased very slightly on the high side (positively biased).

Variance of b [Var (b)]

Omitting terms with X_i^3, X_i^4, etc., in the denominator.

$$Var(b) \cong \frac{\beta^2}{n^2} \Sigma \left[\frac{E(u_i^2)}{5X^2}\right]_i + \frac{1}{n^2} \Sigma \left[\frac{E(v_i^2)}{X^2}\right]_i \tag{B-11}$$

If $E(u_i^2) = \sigma^2_u X_i^2$ and $E(v_i^2) = \sigma^2_v X_i^2$, then

$$Var(b) \cong \frac{5\beta^2\sigma^2_u}{n} + \frac{\sigma^2_v}{n} \tag{B-12}$$

If $\sigma_u = 0.05$ $\sigma_v = 0.10$, $\beta = 0.75$, and $n = 23$, then

$$Var\ (b) = \frac{5(0.75)^2(0.05)^2}{23} + \frac{(0.10)^2}{23}$$

$$\sigma(b) = 0.027.$$

APPENDIX C

COMPARISON OF TWO MEASUREMENT TECHNIQUES

A problem of considerable importance in air pollution technology is the assessment of a measurement method or comparison of one method with another. Two methods can be compared by conducting a designed test program in which both methods are simultaneously used to measure the concentration of a pollutant. In this approach, more than one measurement may be made by a given technique using different operators. In field applications, only one measurement is normally made by each method; thus, it is necessary to make some judgment on the relative precisions of the methods. This approach will be applied here to data from SSI and the dichotomus collocated samplers.

One method of analysis of relative measurement variabilities for the two-instrument case is as follows:[14,18]

$$X_i = \beta_X + p_i + e_{iX} \qquad (C\text{-}1)$$
$$Y_i = \beta_Y + p_i + e_{iY}$$

where

X_i = measurement by instrument X
β_X = bias of instrument X
p_i = true pollution concentration (which varies among samples)
Y_i = measurement by instrument Y
β_Y = bias of instrument Y
e_{iX}, e_{iY} = random errors of measurement by instruments X and Y instruments, respectively.

Two new variables are formed, the sum (s_i) and difference (d_i) of the X_i and Y_i, i.e.,

$$s_i = X_i + Y_i = \beta_X + \beta_Y + 2p_i + e_{iX} + e_{iY}, \qquad (C\text{-}2)$$
$$d_i = X_i - Y_i = \beta_X - \beta_Y + e_{iX} - e_{iY}.$$

The sample correlation of s and d [r(s,d)] is computed and the results applied to compute the t test statistic as follows:[19]

$$t = r(s,d)\sqrt{n-2}/[1 - r^2(s,d)]^{1/2}, \qquad \text{(C-3)}$$

which is distributed as student's t with n-2 degrees of freedom when the population correlation coefficient $\rho(s,d) = 0$.

The population correlation of s and d [$\rho(s,d)$] is

$$\rho(s,d) = (\sigma^2_{eX} - \sigma^2_{eY})/[(4\sigma^2_p + \sigma^2_{eX} + \sigma^2_{eY})(\sigma^2_{eX} + \sigma^2_{eY})]^{1/2}, \qquad \text{(C-4)}$$

which is zero if and only if $\sigma_{eX} = \sigma_{eY}$. Thus, the t test is precisely a test of whether $\sigma_{eX} = \sigma_{eY}$ at the α probability level if the observed t (t_o) is such that

$$-t_{1-\alpha/2} \leq t_o \leq t_{1-\alpha/2},$$

where $t_{1-\alpha/2}$ is the upper $\alpha/2$ probability level of t with n-2 degrees of freedom (df).

We accept $\sigma_{eX} < \sigma_{eY}$ if $t_o < -t_{1-\alpha}$

and that

$$\sigma_{eX} > \sigma_{eY} \text{ if } t_o > t_{1-\alpha}.$$

Example

Table 6 contains the SSI(X) and the TOT (Y) dichotomous sampler data. The sums(s) and differences (d) of the X's and Y's calculated and then their correlation r(s,d),

$$r^2(s,d) = 0.109, \; r(s,d) = 0.33.$$

The t test is given by Equation (C-3) for n = 23 and r(s,d) = 0.33,

$$t = 0.33\sqrt{21}/[1 - 0.109]^{1/2}$$
$$= 1.59 \text{ with } 21 \text{ df},$$

and hence t is not significant at $\alpha = 0.05$ ($t_{0.975(21df)} = 2.08$). We conclude that the two measurement techniques are not significantly different with respect to their precision.

APPENDIX D

MODEL DEVELOPMENT

Let $W_1, \ldots, W_n$ be a random sample from a normal population with mean μ and variance σ^2 and suppose

$$W_i = \ln Y_i, \quad i=1, \ldots, n.$$

Then the expected value of Y_i (EY_i)

$$EY_i = \exp(\mu + \tfrac{1}{2}\sigma^2), \quad i=1, \ldots, n.$$

Because

$$\ln(EY_i) = \mu + \tfrac{1}{2}\sigma^2, \quad i=1, \ldots, n.$$

the minimum variance unbiased estimate of $\ln(EY_i)$ is given by

$$\widehat{\ln(EY_i)} = \hat{\mu} + \tfrac{1}{2}\hat{\sigma}^2, \quad i=1, \ldots, n.$$

From basic theory we have,

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} \ln(Y_i),$$

$$\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n} (\ln Y_i - \hat{\mu})^2.$$

Following Land[20] we take as an estimate for EY_i

$$\widehat{EY}_i = \exp(\hat{\mu} + \tfrac{1}{2}\hat{\sigma}^2).$$

Now suppose that $\ln Y_i$ ($=W_i$) is linearly related to $\ln X_i$, with random error e_i, as follows:

$$W_i = \ln Y_i = \alpha + \beta \ln X_i + e_i$$

$$EW_i = \alpha + \beta \ln X_i = \mu,$$

$$\text{Var } W_i = \text{Var } e_i = \sigma^2, \quad i=1,\ldots, n.$$

If subscripts are suppressed, then

$$\widehat{EY} = \exp(\hat{\alpha} + \hat{\beta} \ln X + \tfrac{1}{2}\hat{\sigma}^2)$$

$$= X^{\hat{\beta}} \exp(\hat{\alpha} + \tfrac{1}{2}\hat{\sigma}^2).$$

Now suppose the observed value of $\hat{\beta}$ is close to unity, so that

$$\hat{\beta} = 1 + \hat{f},\ \hat{f} \ll 1;$$

then

$$\widehat{EY} = (X + \hat{f}\ X\ \ln X)\ \exp\ (\hat{\alpha} + \frac{1}{2}\hat{\sigma}^2)$$

$$\widehat{EY} = \hat{\beta}_1 X + \hat{\beta}_2\ X\ \ln X,$$

where

$$\hat{\beta}_1 = \exp\ (\hat{\alpha} + \frac{1}{2}\hat{\sigma}^2)$$

and

$$\hat{\beta}_2 = \hat{f}\hat{\beta}_1.$$

Now identify TSP with X and one of the IP measurements SSI, TOT, FNE, or CRS with Y, so that

$$\widehat{E(IP)} = \hat{\beta}_1 \cdot TSP + \hat{\beta}_2 \cdot TSP \cdot \log\ TSP.$$